4-3-2020

New Approaches To The Design And Analysis Of Non-Inferiority Clinical Trials

Yulia Sidi

University of Connecticut - Storrs, yulia.sidi@uconn.edu

New Approaches To The Design And Analysis Of Non-Inferiority Clinical Trials

Yulia Sidi, Ph.D.
University of Connecticut, 2020

ABSTRACT

Clinical trials are an essential part of the drug development life cycle. There are different types of clinical trials, and in this dissertation, we focus on non-inferiority (NI) trials. In NI trials the goal is to show that the effectiveness of a new treatment is not considerably worse than of a standard one by an acceptable margin. Although, the new treatment could be slightly less efficacious, it can offer other benefits such as less severe adverse reactions.

Several methodological challenges have been reported regarding the design, analysis and interpretation of NI trials. These include incomplete data analysis, specification of an acceptable margin, and overall benefit of the new non-inferior treatment. Therefore, the aim of this dissertation was to address each of these challenges and provide practical solutions for researchers involved with NI trials.

First, we focus on incomplete data. Specifically, we evaluate how different statistical strategies perform under several NI scenarios and various types of missingness. We provide a set of recommendations for practitioners to use when confronted with incomplete

data to avoid false non-inferiority conclusions. Second, while performing a thorough investigation of proper statistical strategies for incomplete data analysis, we discovered that combination rules of multiply imputed data when inference is done using a Newcombe's method did not exist. As a result, we developed these combination rules. Third, we proposed a new framework that allows for a transparent and objective justification of an acceptable margin. The framework is based on combining results of NI study and clinical experts survey data using multiple imputation (MI). Fourth, we developed a new approach for a comprehensive benefit-risk assessment of a non-inferior treatment. We focus on preference elicitation regarding benefits and risks from a small sample of NI trial participants, and use MI to restore preferences of all study participants.

This dissertation provides an important contribution to the field of Statistics, and drug development. The novel methods and techniques outlined in this dissertation facilitate practitioners involved with NI trails to make more efficient and transparent evaluations of treatment effectiveness.

New Approaches To The Design And Analysis Of Non-Inferiority Clinical Trials

Yulia Sidi

B.A., Hebrew University of Jerusalem, Israel, 2008

M.A., Hebrew University of Jerusalem, Israel, 2011

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2020

APPROVAL PAGE

Doctor of Philosophy Dissertation

New Approaches To The Design And Analysis Of Non-Inferiority Clinical Trials

Presented by

Yulia Sidi, B.A., M.A.

Major Advisor _____

Ofer Harel

Associate Advisor _____

Ming-Hui Chen

Associate Advisor _____

Haim Bar

University of Connecticut

2020

Acknowledgements

I would like to thank my advisor Prof. Ofer Harel for the consistent support and guidance during my dissertation work. I would also like to thank Prof. Nitis Mukhopadhyay, and Prof. Haim Bar who helped me to successfully transition into graduate student life during my first semester at UCONN. Big thanks, also, to Kate McLaughlin, who I extremely enjoyed working with as a TA during the past four years. I would also like to thank Dr. Tim Moore for helping me with the edits of this dissertation.

I wish to express my deepest gratitude to my family, especially my husband Yoni who has been extremely supportive throughout this process. I wouldn't be able to work on this dissertation and complete it without his help. Special thanks to Miki and Dan, whose energy kept me going. I would also like to thank my grandmother, Clara, my mom, Natasha and my aunt Ira. These three women are the reason I appreciate higher education and enjoy studying new things.

Thank you, all the faculty members, staff and fellow students for making this Department of Statistics a perfect place for me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Non-inferiority trials background

Non-inferiority (NI) clinical trials seek to demonstrate that efficacy of a new treatment is not considerably worse than that of a standard treatment [FDA, 2016]. Such minimally acceptable deviation is called margin. The margin is determined using historical data for the standard treatment effect over placebo, and clinical expert opinions regarding the clinically acceptable reduction of that effect. While a portion of a standard treatment effect may be lost by a non-inferior agent, it offers other benefits, such as less severe adverse events, improved drug adherence and/or lower costs [Piaggio et al., 2012]. NI trial design is usually considered when the use of placebo is unethical, as delaying treatment with standard care would cause irreversible health damage or death [ICH, 2000, FDA, 2016].

In the past, NI trials were relatively rare, however, the number of clinical trials using this design increased significantly over time. Suda et al. [2011] evaluated 583 NI trials published between 1989 and 2009, and reported a steadily increasing publication rates. Murthy et al. [2012] perfomed a similar evaluation between 1999 and 2010, which led to

a similar conclusion. Moreover, we searched PubMed in Oct-2019, and found that the number of clinical articles mentioning non-inferiority in title or abstract was 296 in 2010 and reached 713 in 2017 (Figure 1).

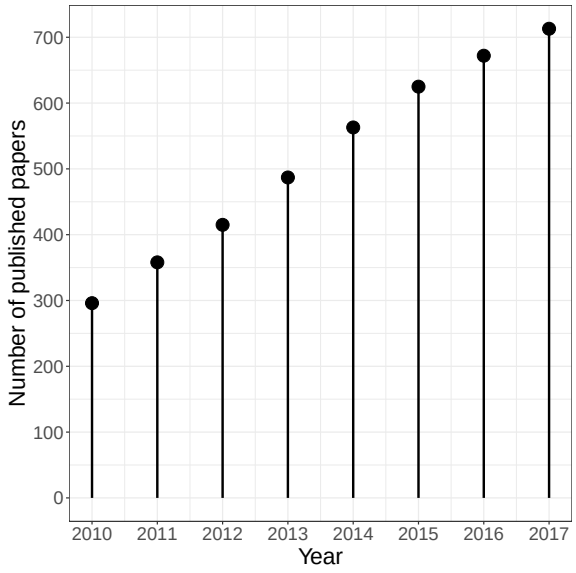

Figure 1: Number of published non-inferiority clinical trials

There are two different approaches to analysis of NI trials: fixed margin method and synthesis method [Rothmann et al., 2016, FDA, 2016]. Both approaches rely on historical data and the opinions of clinical experts, but in a different way. The fixed margin method could be seen as a two step approach. At the first step, one needs to determine the standard treatment effect over placebo (M_1), which is usually done through a meta-analysis of historical data. Then, a clinically acceptable margin (M_2), which has to be strictly lower than M_1 is chosen by clinical experts. The comparison between a standard and a new treatment from an NI trial is done with respect to the pre-defined margin M_2, as if it was a fixed constant. In the synthesis method, the

historical data are "combined" with the NI trial data, so that the new treatment is compared to putative placebo using a clinically acceptable threshold. A recent FDA guideline for NI trials recommends using fixed margin over synthesis method [FDA, 2016]: "Notwithstanding that the interpretation of an NI study is fundamentally a synthesis, we recommend a statistical method, the fixed-margin method, that treats the problem in two separate steps." Following FDA's recommendation, the fixed margin approach was used throughout this dissertation along with a "commonly used" method of 95%-95% confidence interval (CI) [FDA, 2016]. The first 95% CI refers to the standard treatment effect over placebo from historical studies, while the second 95% CI compares the standard and new treatments in the current NI study. To determine NI of the new treatment, the the lower/upper bound of the later CI is compared to M_2.

Recent review articles of published NI trials reveal that a substantial improvement is required for design, analysis and reporting of NI trials [Rehal et al., 2016, Aberegg et al., 2017, Rabe et al., 2018]. Incomplete data analysis, as well as margin justification, were among the issues discussed in these reviews. Moreover, several other authors indicated that there is a need for proper assessment of the overall advantages of a non-inferior treatment over a standard treatment [Garattini et al., 2003, Gladstone and Vach, 2015, Evans and Follmann, 2016]. Following that, we have devoted this dissertation to these topics, with the goal of improving current practices for design and analysis of NI trials.

1.2 Incomplete data analysis in NI trials - current practices

Like most clinical trials, NI trials are prone having incomplete data, which if not properly analyzed might lead to bias results [Little and Rubin, 2014]. The importance of avoiding missing data, and performing appropriate analysis of incomplete data in clinical trials has been extensively discussed [Fleming, 2011, CMPH, 2010, NRC, 2011, Little et al., 2012, Dziura et al., 2013]. However, missing data has received little attention in NI trials. Wiens and Zhao [2007] describe missing data as one of "the biggest obstacles to interpretation of NI studies" and state that more work on this topic is required. Fleming [2008] briefly mentions missing data as a possible source for biased results which might lead to apparent similarities between treatments. Gallo and Chuang-Steiny [2009] provide some considerations in regards to analysis of incomplete data in NI studies.

A common framework for missing data is based on the following missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [Rubin, 1976, Little and Rubin, 2014]. MCAR essentially means that the missing values in the study are completely independent of the data observed or not observed in the study. MAR implies that missing values depend on observed data. MNAR means that missing values depend on unobserved data. MCAR is unlikely to hold in clinical trials [NRC, 2011], therefore, analysis based on this assumption should be avoided, unless it is imposed by design (for example, randomly sampling trial

participants for an ancillary study) [Graham et al., 2006, Little and Rhemtulla, 2013]. In addition to the missingness mechanism, distinctness between data model parameters and parameter involved in generation of missing values plays a central role in incomplete data analysis. For likelihood and Bayesian based inferences, ignorability is characterized by both MAR and distinctness between the parameters mentioned above. As a result, non-ignorability holds when at least one of these two assumptions is violated. For simplicity, we will use MAR/ignorabile and MNAR/non-ignorabile terms interchangeably throughout this dissertation.

Previously, only a handful of simulation studies have been conducted to assess the impact of different analysis strategies on NI trials [Yoo, 2010, Wiens and Zhao, 2007, Lipkovich and Wiens, 2017]. Yoo [2010] conducted a simulation study for longitudinal continuous outcome to evaluate type-I error behavior under different amount and types of missingness. Wiens and Rosenkranz [2013] reported a similar simulation study to that by Yoo [2010]. Both papers concluded that mixed effect repeated measures controlled type-I error under MAR mechanism. Following that, Lipkovich and Wiens [2017] evaluated multiple imputation (MI) [Rubin, 2004] for binary response variables in the NI setting and concluded that MI produces reliable inferences under the MAR assumption. While these simulation studies demonstrate important results, they only consider limited scenarios. Also, dealing with data MNAR remains an unresolved issue.

The lack of deliberation around the missing data problem is also evident in published NI trials. Rehal et al. [2016] reported that over 50% of the published NI trials

they reviewed between 2010-2015 did not mention any imputation methods used in the statistical analysis. Similarly, Rabe et al. [2018] showed that 50% of NI and equivalence articles they reviewed between 2015-2016 used complete case analysis (CCA), a method that is generally known to produce biased results [Little and Rubin, 2014].

In terms of regulatory guidelines, there seems to be lack of consensus around this issue as well [Rehal et al., 2016]. The International Conference on Harmonisation (ICH) guideline for statistical principles for clinical trials [ICH, 1998] mentions missing data issue and states that imputation methods from very simple to complex may be used for analysis of incomplete data. The guideline, however, is very broad, and does not explicitly discuss analysis of incomplete data for NI design. The extension of Consolidated Standards of Reporting Trials (CONSORT) 2010 statement [Piaggio et al., 2012] does not refer to missing data at all. The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) 2013 guideline [Chan et al., 2013] recommends using MI over single imputation and also to perform a sensitivity analysis for missing data. This guideline is however general for any type of clinical trial and does not specifically refer to the NI setting. A recent Food and Drug Administration (FDA) guideline for NI clinical trials [FDA, 2016] includes a very brief statement about missing data: "Imputation of missing data under the inferiority null hypothesis is one possible approach to countering the bias due to attrition". Regulatory guidelines that outline a general framework for handling incomplete data for any type of clinical study are "Guideline on missing data in confirmatory clinical trials" [CMPH, 2010] and "The prevention and treatment of

missing data in clinical trials" [NRC, 2011]. Both of these guidelines provide a comprehensive review of the missing data issue and present methods that could be used to appropriately analyze incomplete data. However, since the scope of these guidelines is general, they don't focus on specific issues that might be relevant for NI trial design. Also, according to the recent ICH E9(R1) guideline, handling of the intercurrent events, such as treatment discontinuation is embedded in the estimand's description [ICH, 2017]. Specifically, the guideline states that occurrence of the intercurrent events in the NI trials using treatment policy strategy might falsely contribute to apparent similarities between the treatment groups, and therefore requires "careful reflection" [ICH, 2017].

Following the above, Chapter 2 of this dissertation focuses on different analysis strategies of incomplete data in NI trials under various missingness structures, as well as diverse set of NI trial scenarios. Specifically, we evaluate the performance of several incomplete data analysis strategies when assessing difference in binomial proportions. While in theory any type of outcome could be assessed for NI treatment comparison (such as means, binomial proportions, survival rates, etc.), binomial proportions were recently reported as the most commonly used outcome in practice. Rabe et al. [2018] reported that (64%) of the articles they assessed had a binomial outcome for primary analysis. In addition, Aberegg et al. [2017] reported that 70% of the studies they evaluated, used absolute risk difference as primary outcome measure. Although an absolute risk differences isn't necessarily based on binomial proportions, it is reasonable to assume that the majority of the studies used it as difference in binomial proportions. According to our own review

of 189 NI clinical studies published on PubMed between June 2017 and May 2018, we found that binomial proportions were used in 71 (38%) studies. The difference between our result and the previously conducted reviews may be due to the somewhat different search criteria: Aberegg et al. [2017] looked at the high impact factor journals, while Rabe et al. [2018] randomly selected papers from a bigger pool of publications. Furthermore, the majority (90%) of the studies with binomial outcomes we reviewed assessed difference in proportions for primary analysis.

There is abundant statistical literature regarding CI construction for difference between binomial proportions and/or testing difference between two binomial proportions. Some earlier work includes well known methods by Clopper and Pearson [1934], Fisher [1935], Barnard [1945], and Pearson [1947], while later methods were developed by Newcombe [1998], Miettinen and Nurminen [1985], Farrington and Manning [1990], Agresti and Caffo [2000] and many others. In line with previously mentioned FDA recommendation to use CI for NI evaluation, as well as, the variety of the methods available for CI construction, we first evaluated what are the most commonly used methods for NI trials in practice. According to our review of 71 NI studies, we found that the most common methods were Newcombe [1998] (NW) method (13%), generalized linear model (GLM) approach (11%), Farrington and Manning [1990] (FM) method (10%), and Wald [1943] method (8%). We also found that 25 (35%) of the reviewed papers didn't explicitly specify what method was used for CI construction. It is reasonable to assume that the majority of these 25 papers used Wald due to its simplicity. Following that, we decided

to use NW, FM and Wald methods in Chapter 2. GLM approach was not considered, because unadjusted GLM using the delta method [Reeve, 2018] to estimate the standard deviation of the difference in proportions has the same form as Wald.

Evaluation of the above methods in NI trials was previously carried out by Dann and Koch [2008]. The authors concluded, that in terms of type-I error, the choice of the method would depend on the allocation ratio between the treatment groups. Similar methods comparison involving Wald CI for difference in binomial proportions for NI trials was done by Li and Chuang-Stein [2006] and by Almendra-Arao [2009]. In addition, Brown and Li [2005] who evaluated performance of several methods for constructing CIs for difference in proportions, which included Wald and NW irrespective of the NI trials setting, concluded that "all the CIs are doing well" when sample size per arm is at least 50. Nevertheless, all of the above evaluations were done for fully observed data. Therefore, comparison between the above methods for incomplete data analysis provides an important addition to the current literature.

For the analysis strategies in Chapter 2, we used best and worst case scenario imputation, CCA and two-stage MI [Shen, 2000, Siddique et al., 2012, 2014]. We provide a thorough explanation for the choice of the strategies in Chapter 2. Briefly, MI is a principled, commonly used approach, which could be used for both ignorable and non-ignorable missingness structures [Rubin, 2004, Sidi and Harel, 2018]. Both, conventional MI and two-stage MI comprise of imputing the incomplete data several times, analyzing each complete dataset using a standard statistical procedure and combing the results

for a final inference [Shen, 2000, Rubin, 2004]. These steps could be easily implemented for Wald and FM methods as shown in Chapter 2. However, it is not the case for NW method, for which there were no proper combination rules found in the literature. Following that, we developed proper combination rules for both conventional and two-stage MI for NW method in Chapter 3.

1.3 Challenges with NI margin choice

As previously mentioned, one of the challenges of NI trials is the choice of clinically acceptable margin (M_2). Although the determination of the margin has been extensively discussed in the literature [Hung et al., 2003, 2005, 2007, Ng, 2008, Hung et al., 2009, Hung and Wang, 2013, Liu et al., 2015], the reasons for choosing a specific margin still remain poorly reported in practice. According to systematic reviews of published NI and equivalence trials, margin justification was mentioned by 45.7%, 23%, 45%, 42.1% and 38% as reported by Wangge et al. [2010], Schiller et al. [2012], Rehal et al. [2016], Althunian et al. [2017] and Rabe et al. [2018] respectively. These findings underline challenges associated with the choice of a margin for NI trials. Obviously, just determination of M_1 is very complex, since historical data carries publication bias and previously observed treatment effect embeds some level of uncertainty. However, even if the standard treatment effect is maintained in the current NI study and the study has assay sensitivity, it is not clear how to choose one number M_2, so that it will be clinically acceptable. A

legitimate question that arises here is the degree of subjectivity of the margin choice. Would it be sufficient to discuss the margin with only one clinical expert? What if an investigator who conducts an NI study reaches out to five clinical experts and they all provide different opinions? How should these opinions be incorporated into the current practices of design and analysis of the study?

Since knowing the "true", objective M_2 would be extremely helpful for design and analysis of NI trials, but the "true" margin cannot be observed, we propose to treat it as missing information. We believe that in order to make proper inferences regarding non-inferiority of the new treatment compared to a standard of care, while minimizing subjectivity of the margin choice, it is imperative to conduct a survey upon clinical experts in this regard. Such survey data can be used to make an informed decision regarding NI of the new treatment. As a result, the reasons for the margin choice could be easier communicated to both regulatory authorities and patients who are seeking alternative treatment options. Moreover, using such survey data regulatory authorities and public health officials will have a better set of tools to justify or disapprove NI of a new treatment.

In Chapter 4, we present a general framework for combining results from a clinical experts survey and NI study. Ideally, the clinical experts survey should consists of a representative sample of clinicians. Obviously, such an assumption could be violated in practice by either surveying a very small number of clinicians, and/or by obtaining opinions of, for instance, more conservative experts. If clinicians conservatism, or lack

of thereof, in respect to a clinical margin is related to other data for the representative sample (professional or demographic characteristics of the clinicians), such data could then be utilized to achieve an objective NI decision. In order to reach this goal, we propose to use MI approach within the above framework.

1.4 Overall benefit of the new non-inferior treatment

As discussed above, to outweigh a decreased effectiveness, the new non-inferior treatment needs to offer advantages over the standard of care [Piaggio et al., 2012, Wangge et al., 2013]. Several authors have recommended formal statistical procedures to simultaneously evaluate efficacy and safety in NI trials [Bristol, 2005, Röhmel et al., 2006, Nishikawa et al., 2009]. Bristol [2005] discussed testing simultaneously non-inferiority of an efficacy endpoint and superiority of a safety endpoint, and recommended using Max Test defined by the authors. Röhmel et al. [2006] considered a problem with two primary endpoints, where it was desired to demonstrate non-inferiority of both endpoints with superiority of at least one of them. The authors suggested using a hierarchal type procedure comprising of the following three steps: 1) test all the endpoints for non-inferiority, 2) show overall superiority, 3) perform one-sided superiority test of each endpoint separately. Nishikawa et al. [2009] proposed the union-intersection test of a

composite endpoint which was defined based on assessing non-inferiority in one endpoint and superiority in another. Also, Gladstone and Vach [2015] introduced a new concept: advantage deficit assessment, where efficacy reduction is compared to gain in safety using the advantage deficit ratio.

A shared feature of the above articles is the consideration of two endpoints. However, during any clinical trial, data on multiple endpoints are being collected and compared between the therapies under evaluation. A question that arises here is: how can benefits and risks from a group of endpoints be evaluated, so that the new treatment is beneficial overall? This question resembles the benefit-risk (BR) topic, which continues to receive a lot of attention [EMEA, 2009, FDA, 2014, Thokala et al., 2016, Marsh et al., 2016, FDA, 2018, Eichler et al., 2009]. According to "Structured approach to benefit-risk assessment in drug regulatory decision-making" draft Prescription Drug User Fee Act (PDUFA) V implementation plan [FDA, 2014], the agency states that for any new treatment to be approved, it needs to show that its benefits outweigh the risks. Moreover, in "Benefit-risk assessment in drug regulatory decision making" draft PDUFA VI implementation plan, the FDA [2018] states not only agency's intention to conduct a structured BR assessment, but also mentions importance of incorporating "patient's voice in drug development and decision-making in the human drug review program". The European Medicines Agency (EMEA) also mentioned the need to shift towards a quantitative BR assessment [Eichler et al., 2009]. Efforts towards implementation of such assessment are evident from "The benefit-risk methodology project" by EMEA [2009], that led to publication of several

working packages [EMEA, 2011, Phillips et al., 2011a,b, 2012, 2014].

Various methods exist for structured BR assessment. Mt-Isa et al. [2016] provides a summary of systematic reviews in this regard. The authors report on quantitative methods that were previously reviewed by the following authors: Guo et al. [2010], Phillips et al. [2011a], Puhan et al. [2012], TORPA [2012], and Mt-Isa et al. [2014]. The multi-criteria decision analysis (MCDA) was reported as the only method, which was both reviewed by each of the above authors, and recommended by two, including EMEA [Phillips et al., 2011a]. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) issued a two-part report for "MCDA Emerging Good Practices Task Force" [Thokala et al., 2016, Marsh et al., 2016]. The ISPOR report described MCDA as a "useful tool" for BR assessment. Moreover, it provided examples of various types of health care decisions that were supported using MCDA [Thokala et al., 2016], which underlines the flexibility and generality of this approach.

MCDA for BR assessment of medicines was first introduced by Mussen et al. [2007]. MCDA allows to combine benefit and risk criteria evidence into one overall assessment measure using scoring and weighting of thereof [Mussen et al., 2007]. The scoring in MCDA transforms each criteria (outcome value) into a common scale, while weighting of the criteria specifies a relative importance of each criterion. For NI trials, such overall assessment measure (MCDA score) could be very useful, since it would allow researchers to claim overall benefit of a new treatment over a standard of care in spite of decreased effectiveness in the primary endpoint.

In Chapter 5, we develop a simple MCDA approach for structured BR assessment of the new non-inferior treatment when compared to the standard of care using patient level data from the NI study. With the growing recognition of the importance of preference elicitation by the patients [Marsh et al., 2017, FDA, 2018] we propose carrying out such elicitation at the beginning of NI trials. Since patients demographic characteristics and baseline disease status are likely to influence patients outcomes, as well as patients preferences, we believe that it would be beneficial to study these within the same trial. The introduction of any questionnaire is likely to increase the burden on the study participants, and investigators, as well as, on the sponsor conducting the study. Therefore, we suggest gathering preferences information only on a random sample of the trial patients, and using MI analysis to create an overall BR assessment. It should be noted that up to now there is limited evidence regarding the actual use of the patient preference in the decision-making process for medical product lifecycle [van Overbeeke et al., 2019].

1.5 Contribution of this dissertation

The main goal of this dissertation is to present innovative methods and practical solutions to issues faced in the design, analysis, and interpretation of NI clinical trials. Each Chapter presented here contributes to this goal. In Chapter 2, we provide a set of recommendations for incomplete data analysis along with a novel approach for analysis

of data under MNAR. This contributes towards better analysis practices of NI trials. In Chapter 3, we introduce MI combination rules for difference in binomial proportions, when the NW method is used. Although we use this method for analysis of NI trials and therefore contribute towards better analysis of such trials, it is a general methodology which is useful for other applications as well. In Chapter 4 we present a new framework for incorporating different clinical experts opinions regarding NI margin into design and analysis of NI trials. While in Chapter 5, we develop a simple BR assessment approach for evaluation of overall benefit of non-inferior treatment. Both Chapter 4 and 5 advance design, analysis and interpretation of NI clinical trials. We hope, that practitioners who are involved with NI trials would find the research done in this dissertation helpful for their own work.

Chapter 2

Incomplete data analysis of NI clinical trials: difference in binomial proportions case

2.1 Background

In this Chapter, we focus on incomplete data analysis for NI clinical trials. We assess the difference in binomial proportions case due to the reasons stated in Chapter 1. We evaluate best and worst-case scenario imputation, CCA and two-stage MI. Best and worst-case scenario imputations and CCA strategies were chosen because these are frequently used in practice for analysis of incomplete NI and equivalence clinical trials data [Rabe et al., 2018].

Both best-case and worst-case scenario strategies were employed only for the MCAR missingness mechanism. It was expected that these two strategies would inflate type-I errors, since they make the two treatment groups more alike, which is anti-conservative in

NI trials. It is well known that while CCA generates unbiased estimates under MCAR, it is generally not the case for MAR [Little and Rubin, 2014]. Conventional MI, on other hand, produces unbiased results under both MCAR and MAR [Little and Rubin, 2014], and therefore is usually recommended over CCA. Despite this, there are still certain conditions under which CCA would result in unbiased estimates under MAR and therefore could be safely used [Bartlett et al., 2015]. The advantage of conventional MI over CCA for NI trials assessing difference in binomial proportions under MAR was previously shown by Lipkovich and Wiens [2017] in terms of unbiasedness and control of the type-I error. The authors, however, did not evaluate cases in which CCA provides unbiased estimates of the treatment effect. Therefore, we explore such conditions here. In addition, conventional MI may result in biased estimates under MNAR unless relevant auxiliary variables are included in the imputation model [Collins et al., 2001, Demirtas and Schafer, 2003]. The inflation of type-I error for NI trials under MNAR, when analyzed with conventional MI was reported by Lipkovich and Wiens [2017]. To resolve the issue of type-I error inflation for NI trials under MNAR, we propose using the two-stage MI procedure described in Section 2.2.

2.2 Methods

2.2.1 Confidence intervals for difference in proportions

Let $Y_{ij} \sim Bernoulli(p_i)$ be an occurrence of a favorable event (such as healing from a disease) for subject j, in a treatment group i. $j = 1 \dots n_i$, where n_i is a sample size of group i and $i = Con,\ Trt$ represents control (or standard), and new treatment respectively. p_i is the true proportion of favorable events in group i.

The hypothesis we are interested in testing is of the following form:

$$H_0 : p_{Con} - p_{Trt} \geq M_2 \ vs \ H_1 : p_{Con} - p_{Trt} < M_2 \,, \tag{2.1}$$

where M_2 represents a pre-defined margin as described in Chapter 1, which is assumed to be positive $M_2 > 0$. H_0 will be rejected at the pre-specified α level if the upper bound of the $100(1-\alpha)\%$ CI for $p_{Con} - p_{Trt}$ is below M_2. As described in Chapter 1, Wald, FM and NW methods for CI construction were used, since these are the most commonly used methods in practice.

Let $\hat{p}_{Con} = \frac{1}{n_{Con}} \sum_{j=1}^{n_{Con}} Y_{Con,j},\ \hat{p}_{Trt} = \frac{1}{n_{Trt}} \sum_{j=1}^{n_{Trt}} Y_{Trt,j}$ be maximum likelihood estimates (MLEs) for $p_{Con},\ p_{Trt}$ respectively, and let $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of a standard normal distribution. The approximate $100(1-\alpha)\%$ CI for $p_{Con} - p_{Trt}$ using the Wald

method has the following form:

$$\hat{p}_{Con} - \hat{p}_{Trt} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_{Trt}(1-\hat{p}_{Trt})}{n_{Trt}} + \frac{\hat{p}_{Con}(1-\hat{p}_{Con})}{n_{Con}}}\,. \tag{2.2}$$

The FM method has a similar form to that of Wald's CI, with the only difference at the variance term estimation, where $\tilde{p}_{Con}, \tilde{p}_{Trt}$ are maximum likelihood estimates of p_{Con}, p_{Trt} respectively under the restriction of the null hypothesis in (2.1) [Farrington and Manning, 1990]:

$$\hat{p}_{Con} - \hat{p}_{Trt} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_{Trt}(1-\tilde{p}_{Trt})}{n_{Trt}} + \frac{\tilde{p}_{Con}(1-\tilde{p}_{Con})}{n_{Con}}}\,. \tag{2.3}$$

Finally, the NW method is based on the Wilson's score method for a single proportion [Wilson, 1927, Newcombe, 1998]. Let LB, UB be a lower and an upper $100(1-\alpha)\%$ CI bounds for $p_{Con} - p_{Trt}$ respectively, defined as:

$$LB = \hat{p}_{Con} - \hat{p}_{Trt} - \sqrt{(\hat{p}_{Con} - l_{Con})^2 + (u_{Trt} - \hat{p}_{Trt})^2}\,, \tag{2.4}$$

$$UB = \hat{p}_{Con} - \hat{p}_{Trt} + \sqrt{(u_{Con} - \hat{p}_{Con})^2 + (\hat{p}_{Trt} - l_{Trt})^2}\,, \tag{2.5}$$

where

$$[l_{Con}, u_{Con}] = \Big(\hat{p}_{Con} + \frac{z^2_{\alpha/2}}{2n_{Con}} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_{Con}(1-\hat{p}_{Con})}{n_{Con}} + \frac{z^2_{\alpha}}{4n^2_{Con}}}\Big)/\Big(1 + \frac{z^2_{\alpha/2}}{n_{Con}}\Big)\,, \tag{2.6}$$

$$[l_{Trt}, u_{Trt}] = \Big(\hat{p}_{Trt} + \frac{z^2_{\alpha/2}}{2n_{Trt}} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_{Trt}(1-\hat{p}_{Trt})}{n_{Trt}} + \frac{z^2_{\alpha}}{4n^2_{Trt}}}\Big) / \Big(1 + \frac{z^2_{\alpha/2}}{n_{Trt}}\Big). \quad (2.7)$$

2.2.2 Multiple imputation

As mentioned in Chapter 1, MI is a principled approach, which could be applied for both ignorbale and non-ignorable missingness processes. When the missingness is non-ignorable, a missingness model needs to be specified. In practice, an exact specification of such a model is difficult, if not impossible, as it relies on a set of unverifiable assumptions. Thus the imputation model could be considered missing, and be multiply imputed together with subject-level data using two-stage MI [Siddique et al., 2012, 2014]. This approach incorporates uncertainty associated with both the imputation model and the imputed subject-level data into the final inference using simple arithmetic combination rules [Shen, 2000, Reiter and Raghunathan, 2007, Siddique et al., 2012, 2014].

We will drop treatment and patient related indexes for the following description of MI procedure for simplicity. Let Y_{com} represent completely observed data, which could be decomposed into $Y_{com} = (Y_{obs}, Y_{mis})$ observed and missing data respectively, and let R be an indicator specifying when the data are observed/missing. Also, let θ be a parameter of interest, and ϕ be a nuisance parameter which characterizes the distribution of the missing data mechanism (R). The goal is to make inferences about $P(Y_{com}|\theta)$, however, because some data are incomplete, a missing data mechanism needs to be considered as well. As a result we have the following joint model:

$$P(Y_{com}, R, \theta, \phi) = P(Y_{com}|\theta)P(R|Y_{com}, \phi)P(\theta, \phi)\,. \tag{2.8}$$

When using MI, the general imputation model is based on the predictive distribution of Y_{mis} given observed data Y_{obs}, and missingess indicator (R) using the following integration of the unknown parameters θ and ϕ as follows:

$$P(Y_{mis}|Y_{obs}, R) = \iint \frac{P(Y_{com}, R, \theta, \phi)}{P(Y_{obs}, R)} d\theta d\phi\,. \tag{2.9}$$

Replacing the numerator in (2.9) by (2.8), (2.9) becomes:

$$P(Y_{mis}|Y_{obs}, R) = \frac{1}{P(Y_{obs}, R)} \iint P(Y_{com}|\theta)P(R|Y_{com}, \phi)P(\theta, \phi) d\theta d\phi\,. \tag{2.10}$$

As described in Chapter 1, under ignorability assumption we have MAR and distinctness as follows:

$$P(R|Y_{com}, \phi) = P(R|Y_{obs}, \phi)\,, \tag{2.11}$$

$$P(\theta, \phi) = P(\theta)P(\phi)\,. \tag{2.12}$$

Substituting (2.11) and (2.12) into (2.10) results in:

$$P(Y_{mis}|Y_{obs}, R) = P(Y_{mis}|Y_{obs})\,. \tag{2.13}$$

Thus, under the ignorability assumption, instead of imputing from $P(Y_{mis}|Y_{obs}, R)$, one needs to impute from $P(Y_{mis}|Y_{obs})$, and "ignore" R. Since, MI commonly implemented using ignorability assumption we will call it "Conventional MI" throughout the dissertation.

However, if the incomplete data follows a MNAR mechanism, then (2.11) does not hold and the ignorability assumption is violated. While the violation of ignorability could also arise from non-distinctness between θ and ϕ, we will assume that ignorability is violated by MNAR throughout this dissertation. In this case, the missingness $P(R|Y_{com}, \phi)$ needs to be modeled. Methods, which were developed specifically for non-ignorable missingness include: selection models [Heckman, 1976, Amemiya, 1984], pattern mixture models [Little, 1993], and shared parameters models [Daniels and Hogan, 2008]. The main difference between these methods is the factorization of the joint distribution of (Y_{com}, R).

In addition, a simple way to generate non-ignorable imputed values (Y^{nonign}) from ignorable imputed values (Y^{ign}) is using the following transformation as presented by Rubin [2004]:

$$Y^{nonign} = a \times Y^{ign} . \tag{2.14}$$

For example, if $a = 1.1$, we assume that conditioning on all other observed information, missing values are still 10% greater than the observed values. Since, in our case the sample space of Y is (0, 1), instead of using (2.14), we will adjust the probability of ignorable imputed probability of event ($\hat{p}^{ign}$), to a non-ingorable imputed probability of event ($\hat{p}^{nonign}$) as follows:

$$\hat{p}^{nonign} = a \times \hat{p}^{ign} . \tag{2.15}$$

As described by Siddique et al. [2012, 2014], in practice it is impossible to know the exact imputation model, and therefore it makes sense to take into account uncertainty associated with the choice of the model. Therefore, following previous work by Siddique et al. [2012, 2014], we suggest specification of a distribution for a, which corresponds to specifying a distribution of the imputation model. Such specification needs to be done by either a study team, or by experts who collect the data.

The imputation model distribution represents the study team's belief regarding the magnitude of the bias in the observed rate in treatment group i, and how confident the team is about this belief. These two values could be seen as the center of the

missingness model distribution (μ_{ai}) and its variance (σ^2_{ai}) respectively. For example, if the team believes the study participants were more likely to drop-out due to lack of efficacy in the new treatment, then the team will anticipate that the observed rate in the new treatment is greater than the actual rate. As a result $\mu_{a,Trt}$ below 1 for the new treatment will be chosen, so that the ignorably imputed rate is closer to it's true value. If, for the same study, the team believes that the observed rate in the control treatment is unbiased, then $\mu_{a,Con} = 1$ would represent such belief. As a result, there is a separate imputation model distribution for each treatment group: $a_{Con} \sim N(\mu_{a,Con}, \sigma_{a,Con})$ and $a_{Trt} \sim N(\mu_{a,Trt}, \sigma_{a,Trt})$ for control and new treatment respectively. We chose to use a normality assumption on the a_i distributions for simplicity, although other distributions can easily replace the normal distribution. After the imputation model distribution is specified, we can randomly draw D models from it. Within each of the imputed models, patient-level data can be imputed L times, resulting in $D \times L$ complete datasets.

Using either conventional MI with L complete datasets or two-stage MI with $D \times L$ complete datasets, each of the complete datasets is then analyzed using a standard statistical method, such as methods presented above. Results from the L, or $D \times L$ analyses are then combined using Rubin's [Rubin, 2004] or Shen's [Shen, 2000] combination rules respectively, as described in the following sections.

Conventional multiple imputation combination rules

Let Q be a quantity of interest, that approximately follows a Normal distribution:

$$(Q - \hat{Q}) \sim N(0, U)\,, \tag{2.16}$$

where $\hat{Q}$ is a complete data statistic estimating Q, and U is a complete data statistic for the variance of $Q - \hat{Q}$. We will further assume, that the dataset is imputed L times, so that $(\hat{Q}^{(l)}, U^{(l)})$ represent the estimate and variance of Q respectively from the l^{th} imputed dataset $(l = 1, \ldots, L)$. Using Rubin's combination rules [Rubin, 2004], the L pairs of estimates are then combined as described below.

Let $\bar{Q}$ be the overall mean of the L estimates:

$$\bar{Q} = \frac{1}{L} \sum_{l=1}^{L} \hat{Q}^{(l)}\,. \tag{2.17}$$

Also let $\bar{U}, B$ be the sources of variability, defined as the overall mean of the associated variance estimates, and between model variance terms respectively. Specifically:

$$\bar{U} = \frac{1}{L} \sum_{l=1}^{L} U^{(l)}\,, \tag{2.18}$$

$$B = \frac{1}{L-1} \sum_{l=1}^{L} (\hat{Q}^{(l)} - \bar{Q})^2\,. \tag{2.19}$$

The overall variance T is then defined as:

$$T = \bar{U} + (1 + \frac{1}{L})B\,. \tag{2.20}$$

The final inferences of the multiply imputed data are based on Student's t distribution:

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_\nu\,, \tag{2.21}$$

where $\nu = (L-1)(1 + \frac{\bar{U}}{B(1+1/L)})^2$.

Two-stage multiple imputation combination rules

The premise specified in (2.16) still holds for the two-stage MI, with the only difference being that, following $D \times L$ imputations, we now have $D \times L$ pairs of estimates $(\hat{Q}^{(d,l)}, U^{(d,l)})$ $(d = 1, \ldots, D,\ l = 1, \ldots L)$ of Q and variance of $\hat{Q} - Q$ respectively. Using Shen's combination rules [Shen, 2000], these pairs of estimates are then combined as described below.

Let $\bar{Q}_2$ be the overall mean of the $D \times L$ estimates:

$$\bar{Q}_2 = \frac{1}{DL}\sum_{d=1}^{D}\sum_{l=1}^{L}\hat{Q}^{(d,l)}\,. \tag{2.22}$$

Also let $\bar{Q}_d$ be the mean of the estimated from the d^{th} model:

$$\bar{Q}_d = \frac{1}{L}\sum_{l=1}^{L}\hat{Q}^{(d,l)}\,. \tag{2.23}$$

Let $\bar{U}_2, W, B_2$ be the three sources of variability, defined as the overall mean of the associated variance estimates, within-model and between model variance terms respectively. Specifically:

$$\bar{U}_2 = \frac{1}{DL}\sum_{d=1}^{D}\sum_{l=1}^{L}U^{(d,l)}\,, \tag{2.24}$$

$$W = \frac{1}{D(L-1)}\sum_{d=1}^{D}\sum_{l=1}^{L}(\hat{Q}^{(d,l)} - \bar{Q}_d)^2\,, \tag{2.25}$$

$$B_2 = \frac{1}{D-1}\sum_{d=1}^{D}(\bar{Q}_d - \bar{Q}_2)^2\,. \tag{2.26}$$

Finally, the total variance of $Q - \hat{Q}$ has the following form:

$$T_2 = \bar{U}_2 + (1+\frac{1}{D})B_2 + (1-\frac{1}{L})W\,. \tag{2.27}$$

The final inferences of the multiply imputed data are based on Student's t distribution:

$$\frac{Q-\bar{Q}_2}{\sqrt{T}_2} \sim t_{\nu_2}\,, \tag{2.28}$$

where $\nu_2^{-1} = \left[\frac{(1+\frac{1}{D})B_2}{T_2}\right]^2 \frac{1}{D-1} + \left[\frac{(1-\frac{1}{L})W}{T_2}\right]^2 \frac{1}{D(L-1)}$.

As stated previously, in this part of the dissertation, two-stage MI was used for MNAR. Therefore, we set $\hat{Q}^{(d,l)} = \hat{p}^{(d,l)}$, where $\hat{p}^{(d,l)}$ is the estimated proportion of difference between control and new treatment from l^{th} imputation and d^{th} model. For Wald and FM, the value of $U^{(d,l)}$ was set to the corresponding variance term used in the method as presented under the square root in (2.2) and (2.3). For NW, $\bar{Q}_2$ for each treatment group was plugged into (2.4 - 2.7). It should be noted, that in Chapter 3 we develop a proper method for combining MI results for the NW method.

2.2.3 Simulation study

Simulation of fully observed data

In total, 30 NI clinical trials scenarios were considered. The p_{Con} values were set to the range between 0.6 and 0.95 by increments of 0.05. The M_2 values were set to: 0.05, 0.075, 0.1, 0.15 and 0.2. All possible combinations of the above margins (M_2) and probabilities (p_{Con}) were used, excluding cases where margin was greater or equal to the corresponding failure rate ($1 - p_{Con}$). A margin equal to the corresponding failure rate would mean that the usage of a new treatment doubles a failure rate of the treated

condition. Therefore, a margin greater or equal to the corresponding failure rate, was redefined as half of the original margin. Due to the high volume of the results, we present here only 9 of the 30 scenarios, which are representative of the rest of the results. In addition, we assumed a one-sided type-I error of 2.5%, power of 90%, and 1:1 group allocation ratio.

Since different methods for comparison of binomial proportions might require different sample sizes [Julious and Owen, 2011], sample sizes were calculated for each method separately using assumptions of the scenarios above. For Wald and FM methods, the sample size calculations were performed by inversion of the corresponding CI formulas [Julious and Owen, 2011], while sample sizes for WN were estimated based on 5000 simulations. As a result the sample size per arm (n) ranged between 98 to 2017 patients (Appendix A, Figure A1).

The outcome variable Y (subscripts are omitted for simplicity) was simulated for each subject using a logistic function of treatment group ($Grp = 0$ for control treatment, $Grp = 1$ for the new treatment) and two continuous baseline covariates ($X_1,\ X_2$) as follows:

$$P(Y = 1) = [1 + e^{-(\alpha_y + \beta_1 * X_{1ij} + \beta_2 * X_{2ij} + \beta_{Grpout} * Grp_{ij})}]^{-1}\,. \tag{2.29}$$

Further details regarding parameters setting in the above model are provided in

(Appendix A, Section A.2). The total number of simulated trials per scenario and method under each hypothesis was set to 10,000 repetitions. It should be noted, that increasing the number of repetition did not alter the results.

Simulation of incomplete data

Let R_{ij} be a missing indicator variable for outcome Y_{ij}, such that $R_{ij} = 1$ indicates that the outcome for patient j in group i is missing while $R_{ij} = 0$ means that the outcome for that patient is observed. Upon a generation of the complete datasets, the missing outcome values were imposed using the following logistic function (subscripts are omitted for simplicity):

$$P(R = 1) = [1 + e^{-(\alpha+\beta_{Grp}*Grp+\beta_Y*Y+\beta_{GrpY}*Grp*Y+\beta_{X_2}*X_2)}]^{-1}. \tag{2.30}$$

Parameters $\beta_{Grp}, \beta_Y, \beta_{GrpY}, \beta_{X_2}$ represent effects of treatment group, outcome, treatment group by outcome interaction and baseline covariate X_2 on missingness respectively. In order to impose a specific missingness mechanism (MCAR, MAR and MNAR), different parameter values were used. The overall drop-out rates were set to $5\%, 10\%, 15\%$ and 20%.

For MCAR, all model parameters but α were set to 0 ($\alpha = -log(\frac{1}{DO} - 1)$, DO is a target drop-out rate). For MAR, β_{X_2} was set to $\beta_{X_2} = 1.5$, while β_{Grp} ranged between -0.9 to 0.9 in order to assess unbalanced levels of drop-out rates of 5-15% between the

treatment groups. MNAR was set up to implement scenarios where dropping out of the study is associated with either lack of efficacy in the new treatment or with overwhelming efficacy in the control treatment, therefore both β_Y, β_{GrpY} were set to non-zero values. These two conditions were considered for MNAR, as both would lead to the observed difference between the treatments appearing smaller than it actually is, which leads to an incorrect study conclusion.

2.2.4 Analysis Strategies for Incomplete Data

As discussed in Section 2.1, the following strategies were considered for the incomplete data analysis: best-case and worst-case scenario imputation, CCA and two-stage MI using multiple imputation chained equations (MICE) [Buuren and Groothuis-Oudshoorn, 2010].

Both best-case and worst-case scenario strategies were employed only for the MCAR missingness mechanism. It was expected that these two strategies would inflate type-I errors, since they make the two treatment groups more alike, which in turn makes it easier to reject the null hypothesis presented in (2.1).

Due to the results provided by Bartlett et al. [2015], it was expected that a CCA strategy would lead to unabised estimates of p_{Con} and p_{Trt} under MAR, as specified below (for simplicity we drop the indexes):

$$P(Y=1|X_1=x_1,X_2=x_2,Grp=i,R=0)= \\ \frac{P(R=0|x_1,x_2,Grp=i,Y=1)}{P(R=0|x_1,x_2,Grp=i)}P(Y=1|x_1,x_2,Grp=i)\,. \tag{2.31}$$

It is easy to see that if $P(R=0|x_1,x_2,Grp=i,Y=1)=P(R=0|x_1,x_2,Grp=i)$, i.e., if missingness of the outcome variable follows MAR process, then:

$$P(Y=1|X_1=x_1,X_2=x_2,Grp=i,R=0)=P(Y=1|x_1,x_2,Grp=i)\,. \tag{2.32}$$

For MNAR missingness process, it was expected that single value imputation methods, or CCA would produce biased results with inflated type-I error rates. Although, conventional MI might produce unbiased estimates when relevant auxiliary variables are used [Collins et al., 2001, Demirtas and Schafer, 2003], our simulation set-up did not address such a situation and therefore we anticipated that conventional MI would not be able to provide unbiased estimates for MNAR. In order to properly analyze the incomplete data that follows such missingness process, we used tthe two-stage MI method described above. Two-stage MI was compared to CCA rather than to conventional MI, due to the fact that both CCA and conventional MI ought to produce biased estimates and because CCA is an easy and dominant approach in clinical trials.

As specified in the previous section, two MNAR situations were simulated: drop-out

due to lack of efficacy in the new treatment and drop-out due to overwhelming efficacy in the control treatment.

For the first situation, it was expected that the observed rate in the new treatment group will be higher than it actually is, while the observed rate in the control group will be unbiased, therefore we specified $a_{Trt} \sim N(\mu_{aTrt}, 0.05)$ where μ_{aTrt} was chosen below 1 and $a_{Con} \sim N(1, 0)$. In contrast, in the second situation, it was expected that the observed rate in the control group was lower than it actually is, while the observed rates in the new treatment will be unbiased, therefore we set $a_{Con} \sim N(\mu_{aCon}, 0.05)$, where μ_{aCon} was chosen above 1 and $a_{Trt} \sim N(1, 0)$.

Similar to Siddique et al. [2014], L was set to 2, and D was set to 100. The multiple imputation of the subject-level data within each imputed missingness model (randomly drawn values of a_{Trt}, a_{Con}) was performed using MICE with the two baseline covariates specified above. We also performed sensitivity analysis for a, by specifying different values for μ_{aTrt}, μ_{aCon} and doubling the standard deviation.

2.2.5 Evaluation Criteria

The Wald, FM and NW performances, along with the analysis strategies used to handle incomplete data, were assessed using empirical type-I error, empirical power, and mean relative bias. Type-I error was estimated by the proportion of trials that reject H_0 in (2.1) out of the trials simulated under $H_0 : p_{Con} = M_2 + p_{Trt}$, and was considered appropriately controlled if it fell within $[0.9\alpha, 1.1\alpha] = [0.0225, 0.0275]$ bounds [Roebruck

and Kühn, 1995, Dann and Koch, 2008]. Power was estimated by the proportion of trials that reject H_0 in (2.1) out of the trials simulated under $H_1 : p_{Con} = p_{Trt}$. A relative bias was defined under $H_0 : p_{Con} = M_2 + p_{Trt}$ as $(\hat{p}_{Con} - \hat{p}_{Trt} - M_2)/M_2$ per repetition. A result was considered unbiased if the mean relative bias fell within $[-0.1, 0.1]$ bounds. The negative bias implies that the new treatment (Trt) is worse than it appears, thus a non-inferiority of the new treatment may be incorrectly inferred.

The simulations presented in this Chapter, as well as in the following Chapters of the dissertation were done using R with reproducible code available on the author's github page.

2.3 Results

2.3.1 Missing completely at random

Table 1 presents empirical type-I errors, and mean relative bias for MCAR data under different study scenarios, as well as empirical type-I errors for fully observed data. Results presented in this table correspond to overall drop-out rate of 20%, as these are representative for lower drop-out rates. Also, since the three CI methods showed very similar results, only the Wald method is presented for MCAR. Columns "Full", "Worst", and "CCA" under "Type-I" in the Table 1 correspond to the empirical type-I error results for fully observed data, incomplete data analyzed using worst-case imputation, and CCA strategies respectively. Also columns "Worst", and "CCA" under "Bias" in the

Table 1 correspond to the mean relative bias results for incomplete data analyzed using worst-case imputation and CCA strategies respectively. For example, the first row of Table 1 corresponds to the scenario in which p_{Con} is 0.65, M_2 is 0.05, the type-I error of the completely observed data is 0.026, while the type-I error for incomplete data analyzed using worst-case imputation is 0.103 with mean relative bias of -0.214, and type-I error for CCA is 0.029 with mean relative bias of -0.019. As can be seen across different scenarios in Table 1, the worst-case scenario imputation strategy produced inflated type-I error rates that were more than double that of the completely observed data, along with significantly biased estimates. In contrast, CCA produced unbiased estimates with type-I errors being either within the pre-specified range of [0.0225, 0.0275] or very close to it.

Table 1: Empirical type-I errors and mean relative bias for MCAR, DO=20%, worst-case imputation scenario and CCA strategies, Wald method

		Type-I			Bias	
p_{Con}	M_2	Full	Worst	CCA	Worst	CCA
0.65	0.05	0.026	0.103	0.029	-0.214	-0.019
0.65	0.10	0.027	0.093	0.028	-0.210	-0.016
0.65	0.15	0.025	0.090	0.026	-0.211	-0.015
0.75	0.05	0.025	0.079	0.026	-0.201	-0.002
0.75	0.10	0.026	0.087	0.029	-0.205	-0.004
0.75	0.15	0.023	0.084	0.025	-0.209	-0.009
0.80	0.15	0.024	0.074	0.026	-0.212	-0.011
0.85	0.05	0.023	0.066	0.024	-0.194	0.008
0.85	0.10	0.028	0.067	0.026	-0.198	0.003

Due to the significant inflation of type-I error for worst-case imputation method, empirical power was calculated for the CCA strategy only. As expected, the power

decreases with higher drop-out rates, dropping to 81.5% (Appendix A, Figure A2). Results for best-case scenario imputation were very similar to the worst-case scenario and therefore are omitted.

2.3.2 Missing at random

Figure 2 presents results from 9 scenarios for empirical type-I errors under MAR with balanced drop-out rates, analyzed using CCA. Each box in the Figure corresponds to a different scenario as specified in the grey title at the top of the box. For instance, the top right box corresponds to the scenario with $M_2 = 0.05$, $p_{Con} = 0.85$, and $n = 1071$. In addition, the dashed lines in each box represent bounds for contorlled type-I error, as specified in Section 2.2, and different dot colors represent different CI construction methods. If a dot lies within the dashed lines, it means that the type-I error is controlled. As can be seen in Figure 2, empirical type-I errors were well controlled in most of the scenarios by the three methods. In addition, this strategy (CCA) resulted in unbiased estimates, while the empirical power went down to 81.7% (Appendix A, Figure A3). For unbalanced drop-out rates, as expected CCA showed slight deviations from the desired level of the type-I error, with maximal empirical type-I error equal to 0.0419 for the overall drop-out rate of 20%, when the drop-out rates between the treatment groups differed by 15% (Appendix A, Figure A4). Nevertheless, the mean relative bias fall within the specified bounds for all of the scenarios, methods, and drop-out rates (results for balanced drop-out rates and overall 20% drop out rates are presented in Appendix

A, Figures A5 and A6).

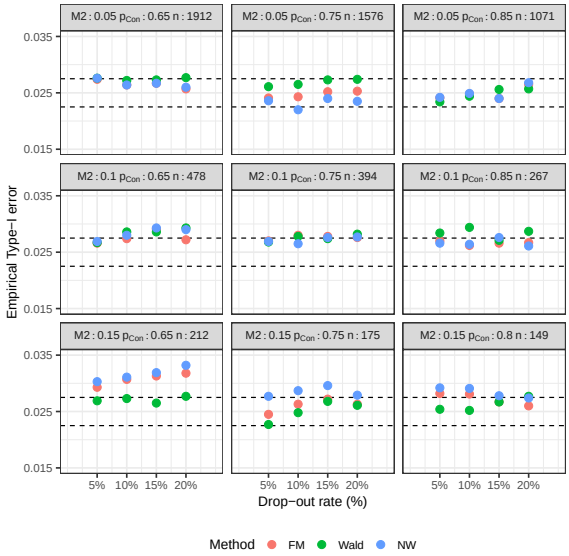

Figure 2: Empirical type-I error CCA strategy for MAR: drop-out rates are balanced between the treatment groups

2.3.3 Missing not at random

Figures 3 and 4 have a similar representation to that of Figure 2 above, and show empirical type-I errors rates for incomplete data under MNAR due to lack of efficacy in the new treatment analyzed using CCA and two-stage MI respectively. As can be seen in Figure 3, empirical type-I errors were seriously inflated when analyzed using CCA. However, as shown in Figure 4, this was not the case for two-stage MI, which produced type-I errors either within the specified bounds or very close to them. In addition, for two-stage MI, the NW method has shown less favorable results compared to the Wald and FM approaches. To further demonstrate advantages of two-stage MI over CCA

for MNAR, we present mean relative bias in Table 2 for drop-out rate of 20% when using Wald method. The corresponding mean relative bias results for the other two methods were similar to Wald and therefore are omitted here. The first two columns of the Table (p_{Con} and M_2) correspond to the scenario assumptions, i.e. event probability in the control treatment and clinically acceptable margin respectively. The later two columns of the Table correspond to mean relative bias following analysis using CCA and two-stage MI, with values below -0.10 indicated biased estimates as specified in Section 2.2. For example, when p_{Con} is 0.65 and M_2 is 0.05, the mean relative bias with CCA strategy is -0.897, while it is only -0.032 with two-stage MI. Overall, as can be seen in Table 2, the CCA strategy produced biased results for all the scenarios, while two-stage MI resulted in unbised estimates. In addition, while the mean relative bias was of a smaller magnitude for lower drop-out rates, CCA still resulted in biased estimates in most cases, while two-stage MI showed unbiased estimates (Appendix A, Table A1). The empirical power based on the two-stage MI was below the desired level of 0.9 with a lowest rate of 65.8% for overall drop-out rate of 20% (Appendix A, Figure A7). This is not surprising due to variability introduced through the MI procedure. Results from MNAR due to overwhelming efficacy in the control treatment were similar in terms of type-I errors, bias and power to the MNAR due to lack of efficacy in the new treatment (Appendix A, Figures A8-A10, and Table A2).

In addition, it should be noted that the only scenario (out of 30), which produced unfavorable results for MNAR when the drop-out rates are due to lack of efficacy in

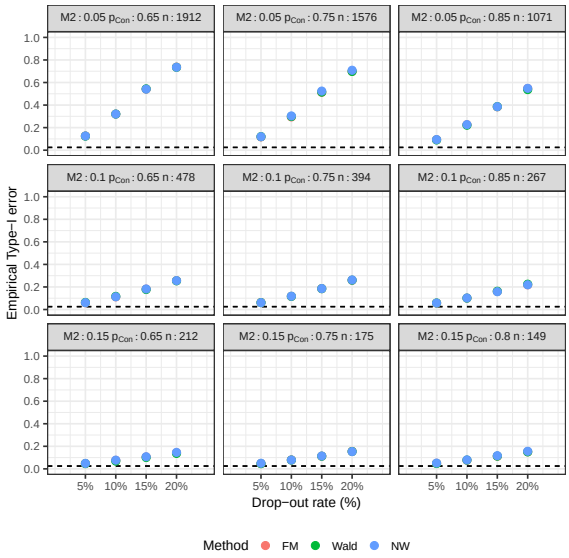

Figure 3: Empirical type-I errors, CCA strategy for MNAR due to lack of efficacy in Trt

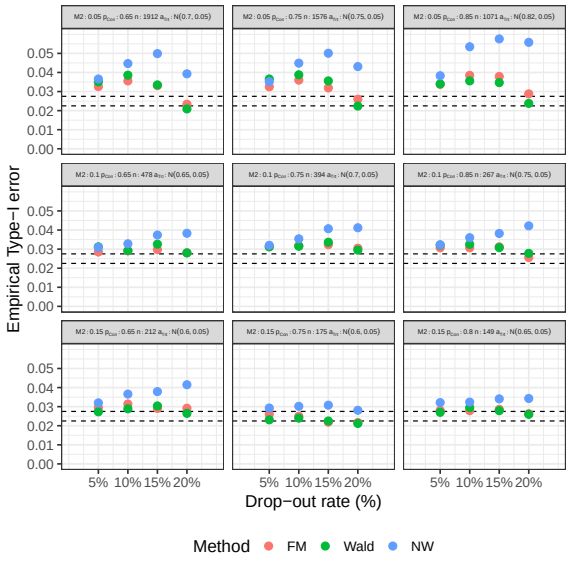

Figure 4: Empirical type-I errors, two-stage MI strategy via MICE for MNAR due to lack of efficacy Trt

the new treatment and the analysis strategy is two-stage MI was the scenario with the highest p_{Con} and lowest M_2 ($p_{Con} = 0.95,\ M_2 = 0.025$). The comparison between the

Table 2: Mean relative bias for MNAR due to lack of efficacy in Trt, DO=20%, CCA and two-stage MI strategies, Wald method

p_{Con}	M_2	CCA	MI
0.65	0.05	-0.897	-0.032
0.65	0.10	-0.453	0.015
0.65	0.15	-0.300	0.022
0.75	0.05	-0.852	-0.038
0.75	0.10	-0.458	0.000
0.75	0.15	-0.319	0.055
0.80	0.15	-0.328	0.019
0.85	0.05	-0.709	-0.068
0.85	0.10	-0.422	-0.001

type-I error rates, as well as mean relative bias between the two analysis strategies for MNAR due to lack of efficacy in Trt using different methods and different drop-out rates is presented in Table 3. For example, the empirical type-I error for 5% drop-out with FM method was 0.086 when analyzed using CCA and 0.044 when analyzed with two-stage MI. Overall, as can be seen in Table 3, the results from two-stage MI performed better than CCA.

In Figure 5, we present a sensitivity analysis for the choice of distribution of imputation models specified by multiplier a_{Trt}. Although type-I error rates are affected by the choice of the imputation model distribution, in all the cases the type-I errors are much smaller than the one observed for CCA strategy (solid black horizontal line).

Table 3: Empirical type-I errors and mean relative bias for scenario with $p_{Con} = 0.95$, $M_2 = 0.025$, MNAR due to lack of efficacy in Trt.

Method	DO	Strategy	Empirical type-I error	Mean relative bias
FM	5%	CCA	0.086	-0.206
FM	5%	MI	0.044	-0.106
Wald	5%	CCA	0.090	-0.206
Wald	5%	MI	0.044	-0.106
NW	5%	CCA	0.086	-0.207
NW	5%	MI	0.054	-0.108
FM	10%	CCA	0.199	-0.392
FM	10%	MI	0.051	-0.195
Wald	10%	CCA	0.209	-0.392
Wald	10%	MI	0.052	-0.195
NW	10%	CCA	0.210	-0.395
NW	10%	MI	0.096	-0.196
FM	15%	CCA	0.359	-0.548
FM	15%	MI	0.039	-0.252
Wald	15%	CCA	0.360	-0.549
Wald	15%	MI	0.042	-0.252
NW	15%	CCA	0.363	-0.552
NW	15%	MI	0.141	-0.256
FM	20%	CCA	0.497	-0.678
FM	20%	MI	0.027	-0.289
Wald	20%	CCA	0.493	-0.676
Wald	20%	MI	0.026	-0.284
NW	20%	CCA	0.499	-0.681
NW	20%	MI	0.168	-0.290

2.4 Conclusion

In this Chapter, we present a thorough simulation study assessing different strategies for analysis of incomplete data when an NI design is employed and the outcome of interest is difference in binomial proportions. We evaluated three commonly used methods for construction of confidence intervals for the difference in binomial proportions: Wald, NW and FM.

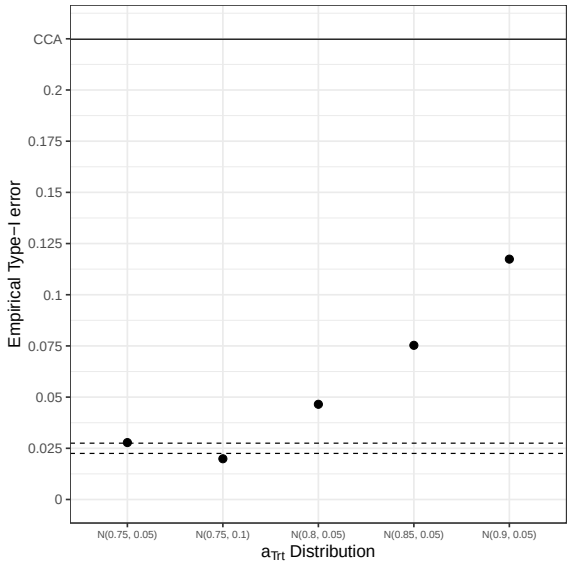

Figure 5: Choice of different distribution parameters for a_{Trt}. Empirical type-I error, two-stage MI strategy via MICE for scenario with $p_{Con} = 0.85, M_2 = 0.1$ for MNAR due to lack of efficacy in Trt using Wald.

We found that both best/worst-case imputation strategies perform poorly even when the incomplete data follows MCAR. This is due to the fact that, by treating incomplete cases similarly for both treatment groups, we make the estimated proportions similar, which leads to erroneous conclusion of NI. According to Rabe et al. [2018] 28% of the reviewed articles that encountered some amount of incomplete data in the primary analysis, used single imputation strategy, including best/worst-case imputation. The simulation results we present here along with the review results reported by Rabe et al. [2018] are concerning. We believe that such an imputation strategy should be abandoned when dealing with NI analysis.

Similar to previous work by Bartlett et al. [2015], we found that CCA performs well when incomplete data follows MAR, and both baseline covariates that affect the

missingness and the corresponding drop-out rates are balanced between treatment arms. In addition, when the drop-outs rates were higher in the new treatment, type-I errors might be inflated, depending on the scenario. Among cases with unbalanced drop-out rates, the highest type-I error rate that was seen is 0.0419% for overall drop-out rate of 20% with 15% higher drop-out in the new treatment. Considering the levels of inflations seen for MNAR and the fact that the 0.0419% rate was reached by a relatively extreme missingness scenario, we believe that CCA could still be considered as a safe choice for MAR incomplete data. It should be noted that if researchers assume that MAR is affected by variables that have different levels between the treatment groups, then a conventional MI strategy is recommended over CCA, as suggested by Lipkovich and Wiens [2017]. The importance of the findings for MAR presented here, is to demonstrate when CCA could be used and what assumption needs to be made in order to have a valid inference.

Importantly, we demonstrate that while CCA performs poorly for incomplete data under MNAR, which is also the case for conventional MI [Lipkovich and Wiens, 2017], two-stage MI strategy produces favorable results. These results are of great importance for practitioners who encounter incomplete data in NI clinical trials. The limitation of this method is the specification of the distribution of the multiple imputation model, or the multiplier. Nevertheless, according to the sensitivity analysis we performed, it is clear that even if the parameters of the multiplier's distribution are shifted, the type-I error rates are still substantially lower than those seen with CCA strategy.

The results of the empirical power were in line with our expectation. In general, empirical power decreased with increasing drop-out rates. In terms of the difference between the analysis methods considered here, we found that in most cases there was no difference between the three. However, when the two-stage MI procedure was used, NW performed worse than Wald and FM. This could be explained by the fact that we used a plug-in method for NW, rather than a proper MI combination rules. In Chapter 3, we present a solution for this issue and show that once implemented the results of NW become closer to the other two methods.

Although, we have looked at a variety of different scenarios, one the the limitations of our work is that it does not cover all possible scenarios. Therefore, before finalizing statistical analysis plan for NI trial, researchers should always consider a specific scenario they are dealing with. Another limitation of our work it that the sample sizes that were considered are moderate to large. We have not evaluated small sample sizes which might require exact methods, such as a method due to Chan [1998], and thus might have different implications when applying MI strategy.

In summary, we recommend employing the following analyses strategies when dealing with incomplete data for NI trials assessing difference in binomial proportions: 1) if the incomplete data follow MAR and it is reasonable to assume that the missingness is caused by balanced baseline covariates only, then CCA could be used, 2) if the data are MAR, but the missingness is caused by other unbalanced variables then following the work by Lipkovich and Wiens [2017] conventional MI should be used, 3) if MNAR is

a more reasonable assumption, then two-stage MI should be used, 4) worst/best-case imputation should be avoided.

The main contribution of this Chapter lies within the above recommendations, which advocate for better analysis practices of NI trials. We believe that these are useful for practitioners who face incomplete data analysis of NI trials that assess difference in binomial proportions.

Chapter 3

Difference between binomial proportions using Newcombe's method with multiple imputation for incomplete data

3.1 Background

As discussed in Chapter 2, if one decides to use the NW method to analyze differences in binomial proportions, and applies an MI procedure due to incomplete data, there are no combination rules to properly estimate the final CI.

The NW method is an extension of the Wilson's score method [Wilson, 1927] for CI construction for one binomial proportion. A similar issue to the above proper combination rules following MI procedure was discussed by Lott and Reiter [2018] for Wilson's method. The authors presented simulation results for MCAR and MAR missingness

mechanisms. Here, we present a proper MI procedure for the NW (MI-NW) method for estimating CI for difference between two proportions, not only for MCAR and MAR but also extend it for MNAR. We compare the NW-MI method developed here with the methods used in Chapter 2, i.e., Wald, FM, and a plug-in NW method (NW-plug).

3.2 Methods

Similar to Chapter 2, $Y_{ij} \sim Bernoulli(p_i)$ is an occurrence of a favorable event for subject j, in a treatment group i ($j = 1 \dots n_i$, $i = Con,\ Trt$), and p_i is the true proportion of favorable events in group i. Also, $\hat{p}_{Con},\ \hat{p}_{Trt}$ correspond to the MLEs of $p_{Con},\ p_{Trt}$ respectively as defined in Chapter 2.

As mentioned above, the NW method is based on the Wilson score method for one proportion [Wilson, 1927]. Specifically, it is assumed that the following large sample approximation holds: $\hat{p}_i|p_i \sim N(p_i, \sqrt{p_i(1-p_i)/n_i})$, which in turn implies that:

$$P(-z_{\alpha/2} < \frac{\hat{p}_i - p_i}{\sqrt{p_i(1-p_i)/n_i}} < z_{\alpha/2}) = 1 - \alpha\,. \tag{3.1}$$

By squaring the term inside the probability in (3.1) and subsequently solving the quadratic equation for p_i, one can get the lower and upper bounds of the CI for p_i as presented in (2.6) and (2.7) for treatment Con and Trt respectively.

The derivation by Lott and Reiter [2018] is based on (3.1). The authors replaced

the $z_{\alpha/2}$, $\hat{p}_i$, p_i, and the denominator $p_i(1-p_i)/n_i$ from (3.1) with t, $\bar{Q}_i$, Q_i, and T_i respectively. $\bar{Q}_i$, Q_i, and T_i come from (2.21) after adding a subscript i for each treatment group, and t is the upper $\alpha/2$ quantile of t_ν distribution in (2.21). As a result the authors presented the following formula for proper combination rules for one binomial proportion using Wilson's method after MI:

$$\frac{2\bar{Q}_i + \frac{t^2}{n_i} + \frac{t^2 r_i}{n_i}}{2(1+\frac{t^2}{n_i}+\frac{t^2 r_i}{n_i})} \pm \sqrt{\frac{(2\bar{Q}_i + \frac{t^2}{n_i} + \frac{t^2 r_i}{n_i})^2}{4(1+\frac{t^2}{n_i}+\frac{t^2 r_i}{n_i})^2} - \frac{\bar{Q}_i^2}{1+\frac{t^2}{n_i}+\frac{t^2 r_i}{n_i}}}, \tag{3.2}$$

where $r_i = \frac{(1+1/L)B_i}{\bar{U}_i}$ and $\bar{U}_i$, and B_i were previously defined in (2.18) and (2.19) respectively.

3.2.1 Multiple imputation Newcombe interval - ignorable missingness

For ignorable missingness, we propose to construct CI for difference between two proportions using a NW-MI method by first performing MI under ignorable assumption as described in Chapter 2 for each group separately, and thus obtaining values $\bar{Q}_{Trt}, r_{Trt}, \bar{Q}_{Con}$, r_{Con}, which can then be used in (3.2) to obtain lower and upper bounds (ql_i, qu_i respectively) for each MI estimated proportion for p_{Con} and p_{Trt}. Similar to the original Newcombe's method for completely observed data [Newcombe, 1998], the resulting

$ql_{Trt}, qu_{Trt}, ql_{Con}, qu_{Con}$ are then used in (2.4) and (2.5) as $l_{Trt}, u_{Trt}, l_{Con}, u_{Con}$ respectively along with $\hat{p}_i = \bar{Q}_i$ to estimate the CI for difference between two proportions $p_{Con} - p_{Trt}$ for the NW-MI method.

3.2.2 Multiple imputation Newcombe interval - nonignorable missingness

For non-ignorable missingness, we propose to use two-stage MI described in Chapter 2. Similar to the previous section, we perform two-stage MI for each group separately and obtain CIs for p_{Con} and p_{Trt}. The lower and upper bounds of these CIs are then used to construct a CI for the difference between proportions.

Let's start by first introducing a properly constructed CI for one proportion using the Wilson method with two-stage MI. It should be noted that our derivations are closely related to, and use similar notation, as in Lott and Reiter [2018]. For simplicity we drop subscript i in the following derivations. Based on the result of combination rules of two-stage MI in (2.28), a $(1-\alpha)100\%$ CI for Q is defined as:

$$P(-t_2 \leq \frac{Q - \bar{Q}_2}{\sqrt{T_2}} \leq t_2) = 1 - \alpha\,, \tag{3.3}$$

where t_2 is an upper $\alpha/2$ quantile of t_{ν_2} distribution in (2.28). Next we square the inside probability terms in (3.3), and substitute T_2 as defined in (2.27). As a result we obtain

the following inequality:

$$\frac{(Q-\bar{Q}_2)^2}{\bar{U}_2+(1+\frac{1}{D})B_2+(1-\frac{1}{L})W} \leq t_2^2\,. \tag{3.4}$$

Then by using the following: $r_2 = \frac{(1+1/D)B_2}{\bar{U}_2}$ and $s = \frac{(1-1/L)W}{\bar{U}_2}$, we have:

$$\frac{(Q-\bar{Q}_2)^2}{\bar{U}_2(1+r_2+s)} \leq t_2^2\,. \tag{3.5}$$

According to Rubin [2004] it is reasonable to assume that $\bar{U}_2 \approx U$. Since in the one proportion binomial case we have $U = p(1-p)/n = Q(1-Q)/n$, (3.5) becomes:

$$\frac{(Q-\bar{Q}_2)^2}{(Q(1-Q)/n)(1+r_2+s)} \leq t_2^2\,. \tag{3.6}$$

By solving (3.6) for Q we get the following lower and upper bounds (q_2l, q_2u respectively) for p:

$$\frac{2\bar{Q}_2+\frac{t_2^2}{n}+\frac{t_2^2 r_2}{n}+\frac{t_2^2 s}{n}}{2(1+\frac{t_2^2}{n}+\frac{t_2^2 r_2}{n}+\frac{t_2^2 s}{n})} \pm \sqrt{\frac{(2\bar{Q}_2+\frac{t_2^2}{n}+\frac{t_2^2 r_2}{n}+\frac{t_2^2 s}{n})^2}{4(1+\frac{t_2^2}{n}+\frac{t_2^2 r_2}{n}+\frac{t_2^2 s}{n})^2} - \frac{\bar{Q}_2^2}{1+\frac{t_2^2}{n}+\frac{t_2^2 r_2}{n}+\frac{t_2^2 s}{n}}}\,. \tag{3.7}$$

As a result, (3.7) represents a proper CI for one proportion using the Wilson method with two-stage MI. Based on this equation, one can construct a CI for NW-MI method for difference between two proportions using (2.4) - (2.8). Specifically, let's assume two-stage MI is used to estimate CIs for p_{Con} and p_{Trt} separately using (3.7). Thus, we obtain lower and upper CI bounds for p_{Con} ($q_2 l_{Con}$ and $q_2 u_{Con}$ respectively) and p_{Trt} ($q_2 l_{Trt}$ and $q_2 u_{Trt}$ respectively). The values $q_2 l_{Trt}, q_2 u_{Trt}, q_2 l_{Con}, q_2 u_{Con}$ are then used as $l_{Trt}, u_{Trt}, l_{Con}, u_{Con}$ in (2.4) and (2.5), respectively to obtain lower (LB) and upper bound (UB) of CI for difference between two proportions $p_{Con} - p_{Trt}$.

It should be noted that for NW, MI prefix indicates that a multiple imputation procedure based on either ignorable or non-ignorable assumption was implemented.

3.2.3 Simulation studies

In order to assess the performance of NW-MI method developed here and compare it to the other three methods used in Chapter 2 we used simulation studies with the following set-up: $p_{Con} \in \{0.65, 0.9\}$, $p_{Trt} = p_{Con} - M_2$, $M_2 \in \{0.025, 0.1\}$, $n_i \in \{100, 500\}$, and it was assumed that $n_{Trt} = n_{Con}$. The values of p_{Con} and M_2 were chosen to represent both mid-range and high value probabilities with small and moderate differences between the two proportions. The sample size values where chosen to assess the impact of moderate and large samples per group. In addition, drop-out (DO) rates were assumed to be balanced between treatment groups and were set to induce both low (10%) and moderate (30%) rates. As a result, we had 16 different scenarios for all possible combinations of

the above specifications. The α level was set to 5%. Simulations were repeated 10,000 times for each scenario. It should be noted, that increasing the number of repetition did not alter the results.

Let R_{ij} be a missingness indicator with $R_{ij} = 1$, when Y_{ij} is missing and $R_{ij} = 0$, when Y_{ij} is observed. In the following sections we explain how MCAR, MAR and MNAR were specified and analyzed.

MCAR - simulation and analysis

MCAR missingness structure was imposed by randomly masking the values of Y to achieve $P(R_{ij} = 1) = DO$. The incomplete data were multiply imputed based on the ignorability assumption. In other words, it was sufficient to use only observed values of Y_{ij} in order to determine the distribution of the incomplete values of Y_{ij}, and thus properly impute them. It should be noted that in our case the determination of the distribution of incomplete values of Y_{ij} essentially means specifying $p_i^* = P(Y_{ij} = 1|R_{ij} = 1)$ correctly, so we can impute these values based on $Bernoulli(p_i^*)$ distribution. We assumed a non-informative prior for $p_i \sim U(0, 1)$, so that the posterior distribution of p_i was $Beta(a_i, b_i)$, where a_i is number of ones in group i plus one and b_i is a number of observed zeros in group i plus one. The value of p_i^* was then randomly drawn from $Beta(a_i, b_i)$, and was used to determine the values of Y_{ij} for incomplete observations using $Bernoulli(p_i^*)$ distribution. Within each imputation, we used 1000 iterations of Gibbs sampling to achieve convergence to a stable stationary distribution of p_i^*. In total

we repeated this procedure $L = 10$ times, which resulted in 10 complete datasets, similar to the number of imputations used by Lott and Reiter [2018]. As part of the sensitivity analysis, we increased the number of imputations to 20, which did not alter the results.

Following the MI procedure, CI's for different methods were constructed as described above for NW-MI and in Chapter 2 for the other three methods.

MAR - simulation and analysis

In order to impose MAR, an additional categorical variable X was defined. Similar to Lott and Reiter [2018], we used two types of MAR, first we assumed strong association between X and Y, specified as follows (subscripts are dropped for simplicity): $P(X = 1|Y = 1) = 0.2,\ P(X = 1|Y = 0) = 0.6$. Then we assumed independence between X and Y by specifying: $P(X = 1|Y = 1) = P(X = 1|Y = 0) = 0.6$. Due to the ignorable missingness imposed here, MAR was imputed using the same procedure as describe for MCAR for each value of X separately. Also, the construction of CIs was done similarly to those done for MCAR imputed data.

3.2.4 MNAR - simulation and analysis

For MNAR, missingness was specified by making probability of missingness depend on the outcome values for group Con as following:

$$P(R_{Con,j}=1|Y_{Con,j}=0)=\frac{P(R_{Con,j}=1)-P(R_{Con,j}=1|Y_{Con,j}=1)P(Y_{Con,j}=1)}{P(Y_{Con,j}=0)},$$

where $P(R_{Con,j}=1|Y_{Con,j}=0)$ was set to be smaller than $P(R_{Con,j}=1|Y_{Con,j}=1)$, and $P(R_{Con,j}=1)=DO$. The probability of missingness for group Trt was set to be independent of the Y values, i.e. $P(R_{Trt,j}=1)=DO$. As a result, the observed difference between the estimated proportions would appear to be smaller than it actually is.

In the situation presented here, it is clear that MI using the ignorable assumption would not result in proper inferences. This is due to the fact that the observed difference between the proportions would be biased as described above, whereas the unobserved information is the source of such bias. To demonstrate this we performed MI using both ignorable and non-ignorable assumptions. For the ignorable assumption, we simply followed the procedure described in Section 3.2.3, since the missingness was generated only using values of Y. For non-ignorability, we used a two-stage MI procedure as follows.

As presented in (2.15), we can adjust an event probability estimated under ignorable assumption ($\hat{p}_i^{ign}$) by using some constant multiplier a, so that $\hat{p}_i^{nonign}$ will represent probability estimated under non-ignorable assumption. The incomplete data is again imputed for each group separately, therefore we specify a_i as multiplier for group i. In

general, specifying different multipliers per group separately allows us to assume different missingness models for the groups, a situation that is not uncommon in practice. First we set a distribution for a_i as $a_i \sim N(\mu_i, \sigma_i)$. Since missingness for group Trt is essentially ignorable we specify $a_{Trt} \sim N(1, 0)$. For group Con, based on (2.15) μ_{Con} can be defined as a link between a probability of a favorable event based on observed and non-observed values as (the subscripts are omitted for simplicity):

$$P(Y = 1|R = 1) = \mu_{Con} \times P(Y = 1|R = 0)\,.$$

Moreover, it is easy to see that:

$$P(Y = 1|R = 1) = \frac{P(Y = 1) - P(Y = 1|R = 0)P(R = 0)}{P(R = 1)}\,, \tag{3.8}$$

and

$$P(Y = 1) = \frac{P(R = 0)P(Y = 1|R = 0)}{P(R = 0)}\,. \tag{3.9}$$

inserting (3.9) into (3.8) gives us:

$$\mu_{Con} = \frac{P(R_{ij} = 0)}{P(R_{ij} = 1)P(R_{ij} = 0|Y_{ij} = 1)} - \frac{P(R_{ij} = 0)}{P(R_{ij} = 1)}\,. \tag{3.10}$$

The value of σ_{Con} ranged between 0.03 to 0.26 and was calibrated through simulation of the 10,000 repeated samples that were generated for MNAR missingness process. Specifically, within each repeated sample, we calculated the expression that appears in (3.10), and σ_{Con} was set to be a standard deviation for these 10,000 values within each scenario.

The values of a_i's were then randomly drawn D times from the distribution specified above. In addition, we estimated $\hat{p}_i^{ign}$ L times using the MI procedure presented in Section 3.2.3, so that $\hat{p}_i^{ign} = p_i^*$. Then, L values of $\hat{p}_i^{ign}$ were multiplied by D values of a_i's, so that we received $D \times L$ $\hat{p}_i^{nonign}$ values. Incomplete values of Y_{ij} were then determined by $Bernoulli(\hat{p}_i^{nonign})$ distribution. As a result we received $D \times L$ complete datasets and used two-stage MI combination rules described in Chapter 2 to summarize them. Similar, to Siddique et al. [2012], two-stage MI for MNAR was performed using $D = 100, L = 2$.

3.2.5 Evaluation criteria

The initial evaluation criteria included coverage probability, average interval width, percent of cases where the interval fell outside the range of $[-1, 1]$, and percent of cases with zero width. Since, after completing all the simulations, we did not encounter any cases where the interval either fell outside $[-1, 1]$ or had zero width intervals, these evaluation criteria are not be presented. A procedure was considered as favorable if it's coverage probability achieved the desired α level. Among favorable procedures, a procedure with

the shortest CI width would be considered more advantageous. Coverage probability was calculated as the number of times the true difference M_2 fell inside the 95% CI divided by 10,000 repetitions.

3.3 Results

Table 4 presents coverage probabilities for fully observed data for each simulation scenario and method. For example, for $p_{Con} = 0.65$, $M_2 = 0.025$ and $n = 100$, the coverage probability for Wald was 0.9476. As can be seen in Table 4, the coverage probability of the fully observed data reached the desired rate of 95% for all the methods. Table 5 presents average widths of the 95% CIs for fully observed data for each simulation scenario and method. For example, for $p_{Con} = 0.65$, $M_2 = 0.025$ and $n = 100$, the average width for Wald was 0.265. As can be seen in Table 5, the average width of the 95% CIs were similar between the methods for the fully observed data.

Table 4: Coverage probability for fully observed data

p_{Con}	M_2	n	Wald	FM	NW
0.65	0.025	100	0.9476	0.9502	0.9502
0.90	0.025	100	0.9455	0.9483	0.9541
0.65	0.10	100	0.9475	0.9489	0.9489
0.90	0.10	100	0.9478	0.9481	0.9513
0.65	0.025	500	0.9526	0.9530	0.9528
0.90	0.025	500	0.9501	0.9507	0.9515
0.65	0.10	500	0.9478	0.9495	0.9495
0.90	0.10	500	0.9479	0.9482	0.9488

Figure 6 shows coverage probabilities for all the scenarios and methods under MCAR.

Table 5: Average width of 95% CI for fully observed data

p_{Con}	M_2	n	Wald	FM	NW
0.65	0.025	100	0.265	0.266	0.261
0.90	0.025	100	0.174	0.174	0.180
0.65	0.100	100	0.269	0.270	0.264
0.90	0.100	100	0.194	0.195	0.198
0.65	0.025	500	0.119	0.119	0.119
0.90	0.025	500	0.078	0.078	0.079
0.65	0.100	500	0.121	0.121	0.120
0.90	0.100	500	0.087	0.088	0.088

The dashed line on the Figure represents the desired coverage probability of 0.95, and the dotted line represents coverage probability of 0.90. The drop-out rates correspond to different colors, and sample size is represented by different shapes. A method is considered to perform well, if its coverage probability is on the dashed line or close to it. As can be seen on Figure 6, NW-MI, Wald and FM show coverage rates close to the desired level. Importantly, the three methods outperform NW-plug, which has coverage rates between 89.2% to 90.4% for drop-out rate of 30%, and between 93.3% to 94.3% for drop-out rate of 10%. Average CI widths for MCAR are presented in Figure 7, as can be seen NW-plug had a shorter CI width than the other methods across all the scenarios.

Figure 8 has a similar presentation as Figure 6, and shows coverage probabilities for MAR with highly correlated X and Y. As can be seen in Figure 8, NW-plug showed coverage rates below 90% for drop-out rates of 30% for most of the scenarios, while the coverage rates for drop-out of 10% were between 92.8% and 94.1%. In contrast, NW-MI, Wald and FM demonstrated coverage probabilities which either achieved or were above

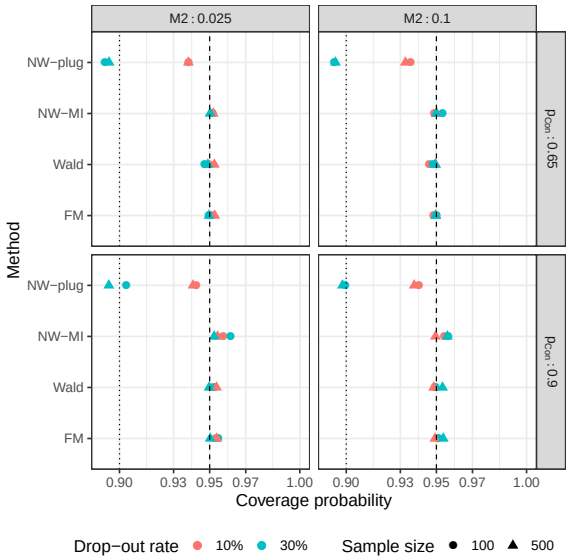

Figure 6: Coverage probability for MCAR (Dashed line represents the desired coverage probability of .95, dotted line represents covarage probability of .90.)

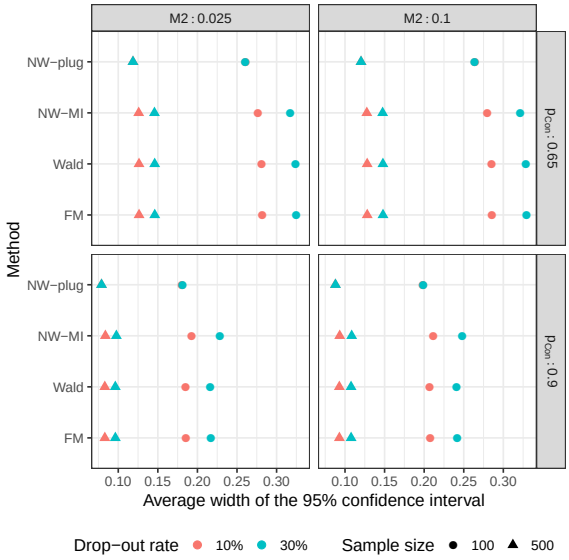

Figure 7: Average width of 95% confidence intervals for MCAR

the desired level of 95%. Average CI widths for MAR with highly correlated X and Y are presented in Figure 9. As can be seen in Figure 9, similarly to MCAR, the average CI widths were shorter with MI-plug than with the other methods. Results for MAR

with independent X and Y are similar to the results for MAR with highly correlated X and Y, and therefore are presented in Appendix B (Figures B1, B2).

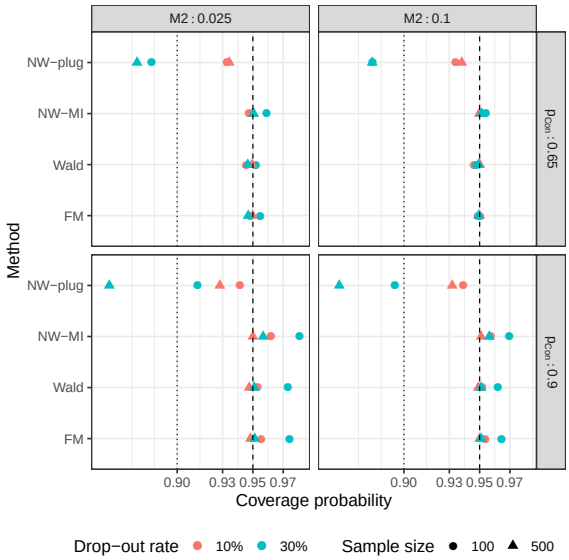

Figure 8: Coverage probability for MAR with highly correlated X and Y (Dashed line represents the desired coverage probability of .95, dotted line represents coverage probability of .90.)

As stated previously, MNAR data were analyzed using the non-ignorable assumption. Figure 10 has similar presentation to that of Figure 6, and shows coverage probabilities for MNAR. As can be seen in Figure 10, the coverage rates for the drop-out of 10% were at or close to the desired level of 95% for NW-MI, Wald, and FM methods, while these were between 92.7% and 94.3% for NW-plug. For the 30% drop-out, the coverage probability ranged between 92% and 94.5% for FM, between 92% and 94.2% for MI-Wald, between 91.3% and 93.9% for MI-NW, and between 84.5% and 89.4% for MI-plug (Figure 10). Average CI widths results were again similar to the ones produced for previous missingness mechanisms (Appendix B, Figure B3).

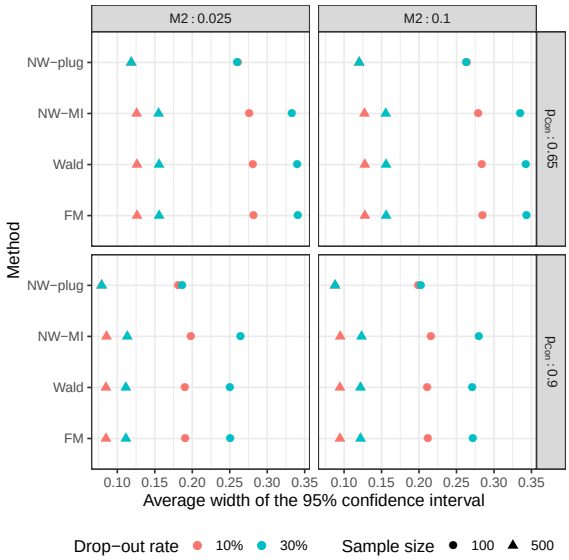

Figure 9: Average width of 95% confidence intervals for MAR with highly correlated X and Y

In order to demonstrate advantages of non-ignorable NW-MI over the ignorable one under MNAR, we present coverage rates of these approaches for all the simulation scenarios and drop-out rates in Table 6. In addition, the last column in the Table shows differences between the coverage probabilities obtained from non-ignorable MI and the coverage probabilities obtained from ignorable MI, so that a positive difference indicates a higher coverage for non-ignorable MI. For example, when $p_{Con} = 0.65,\ M_2 = 0.10,\ n = 500,\ DO = 0.30$, the coverage rate with non-ignorable MI is 74.2% higher than the corresponding coverage rate with ignorable MI. As can be seen in Table 6, the coverage rates obtained from non-ignorable MI were closer to the nominal coverage of 95% than those based on ignorable MI for all scenarios.

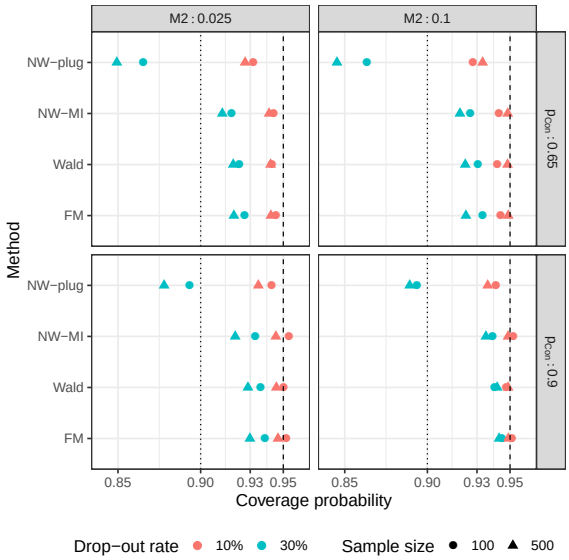

Figure 10: Coverage probability for MNAR analyzed based on non-ingorable assumption (Dashed line represents the desired coverage probability of .95, dotted line represents coverage probability of .90.)

Table 6: Comparison of coverage probabilities estimated under non-ignorability and ignorability assumption for MNAR using NW-MI method

p_{Con}	M_2	n	DO	CP non-ignorable MI	CP ignorable MI	CP difference
0.65	0.025	100	0.10	0.9442	0.9330	0.0112
0.65	0.025	100	0.30	0.9187	0.7451	0.1736
0.65	0.025	500	0.10	0.9414	0.8777	0.0637
0.65	0.025	500	0.30	0.9131	0.1640	0.7491
0.65	0.10	100	0.10	0.9432	0.9317	0.0115
0.65	0.10	100	0.30	0.9259	0.7449	0.1810
0.65	0.10	500	0.10	0.9486	0.8815	0.0671
0.65	0.10	500	0.30	0.9197	0.1776	0.7421
0.90	0.025	100	0.10	0.9535	0.9530	0.0005
0.90	0.025	100	0.30	0.9331	0.9263	0.0068
0.90	0.025	500	0.10	0.9455	0.9358	0.0097
0.90	0.025	500	0.30	0.9209	0.7680	0.1529
0.90	0.10	100	0.10	0.9520	0.9505	0.0015
0.90	0.10	100	0.30	0.9394	0.9253	0.0141
0.90	0.10	500	0.10	0.9488	0.9418	0.0070
0.90	0.10	500	0.30	0.9354	0.7956	0.1398

3.4 Conclusion

In this Chapter, we extended the previously developed proper combination rules for estimating CI for one binomial proportion using the Wilson method with MI [Lott and Reiter, 2018] to estimate CIs for difference between two proportions. Moreover, we developed a proper two-stage MI procedure for constructing CIs using the NW method when the incomplete data follows a MNAR missingness mechanism. We compared the performance of our method to NW-plug, Wald and FM methods in terms of coverage probability and CI width using several simulation scenarios. We showed that the NW-plug method had coverage rates below the desired rate, and lower than the other three methods. The results for NW-MI, Wald and FM were comparable. For MCAR and MAR, we showed that the coverage probabilities for NW-MI, Wald, and FM were at the desired level of 95% for $p_C = 0.6$. Moreover, for $p_C = 0.9$, we observed that for both MCAR and MAR, the coverage probabilities for NW-MI, Wald and FM were either at or above 95%. Importantly, we showed the advantage of using two-stage MI over a conventional MI in terms of the coverage rates. In terms of the average width, NW-plug showed the shortest CIs, while the other three methods had comparable average CI widths.

As a result of the above evaluation, NW-MI, Wald and FM are recommended over NW-plug. Following this, we implemented NW-MI for one of the scenarios from Chapter 2 ($M_2 = 0.05, p_{Con} = 0.85, n = 1071$) for MNAR due to lack of efficacy in Trt. As can

be seen in Figure 11 the results of NW-MI are now comparable with Wald and FM as expected.

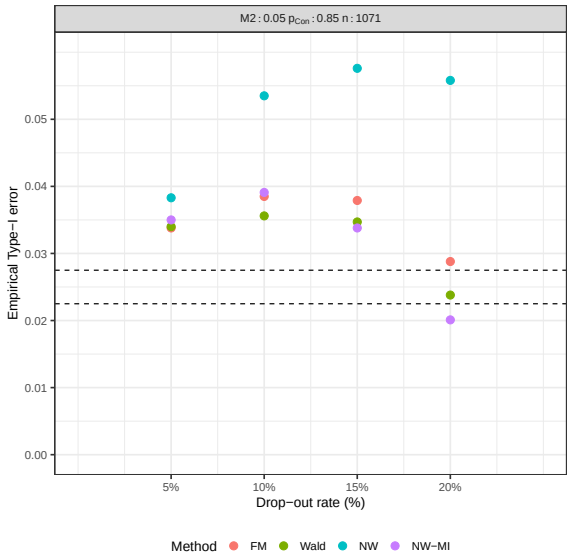

Figure 11: Empirical type-I errors, two-stage MI strategy via MICE for MNAR due to lack of efficacy in Trt, following NW-MI implementation

The limitation of our evaluation is that we could only assess a set of scenarios, which although are representative of many applied research problems, do not cover all possible scenarios. In addition, we did not assess performance of the methods presented here for small sample sizes. The reason for that is that all the methods we used here are based on Normal approximation, and in general are known to perform well for completely observed data for sample sizes above 50 [Brown and Li, 2005]. Nevertheless, for smaller sample sizes exact CI construction methods, e.g. Clopperr-Pearson [Clopper and Pearson, 1934] could be more favorable, however the exact methods are outside of the scope of this work.

We believe that the new methods presented here could be very useful in practice.

First of all, in many areas of applied research the outcome of interest is the difference between two proportions. Second, many studies encounter incomplete observations and have to make inferences given the data they observe. While it is impossible to test the incomplete data for ignorable vs non-ignorable missingness assumption, the ignorability assumption needs to be explicitly stated and consequently translated into a proper type of analysis [Sidi and Harel, 2018]. The contribution of the Chapter is two-fold: first we extend on the previous research conducted by Lott and Reiter [2018], second we propose a relatively simple method which can be used for non-ignorable missingness.

Chapter 4

Non-inferiority clinical trials: treating margin as missing information

4.1 Background

As described in Chapter 1, the choice of clinically acceptable margin (M_2) continues to be a major issue for the design and interpretation of NI trials. The fixed margin approach described in Chapter 1 is illustrated in Figure 12. The upper part of this Figure presents a historical comparison of the standard treatment to placebo to determine the value of M_1, which is the entire effect of the standard treatment over placebo. Following that, M_2 is chosen by clinical experts, its value is presented in the bottom part of the Figure. It should be noted that M_2 is lower than M_1. In order for the new treatment to be non-inferior, the CI in NI trial needs to be below M_2. In addition, several scenarios for the NI trial comparing the standard treatment to the new test treatment are presented

in the bottom part of Figure 12. In those scenarios NI is concluded for the CIs with blue point estimates, while the CI with a red point estimate corresponds to inferiority of the new test treatment. Since different values of M_2 can lead to different conclusions, a proper determination of the clinically acceptable margin is essential for NI trials.

In this Chapter, we propose to treat M_2 as missing information. We propose a survey among clinical experts in order to determine the objective value of the margin, as well as variability associated with it. Following our novel framework, the information obtained from such a survey can then be combined with the NI trial results, which would lead to an objective decision regarding the new potentially non-inferior treatment.

We mentioned in Chapter 1, that while the goal is to survey a representative sample of clinicians, this might be hard to achieve in practice. Therefore, if clinical experts' opinions regarding M_2 are influenced by the experts' professional or demographic characteristics, surveying a small number of clinicians about M_2, while obtaining general characteristics for representative sample is sufficient when utilizing MI.

4.2 Methods

Similar to Chapters 2 and 3, $Y_{ij} \sim Bernoulli(p_i)$ is an occurrence of a favorable event for subject j, in a treatment group i ($j = 1 \dots n_i$, $i = Con,\ Trt$), and p_i is the true proportion of favorable events in group i. Also, $\hat{p}_{Con},\ \hat{p}_{Trt}$ correspond to the MLEs of $p_{Con},\ p_{Trt}$ respectively as defined in Chapters 2 and 3.

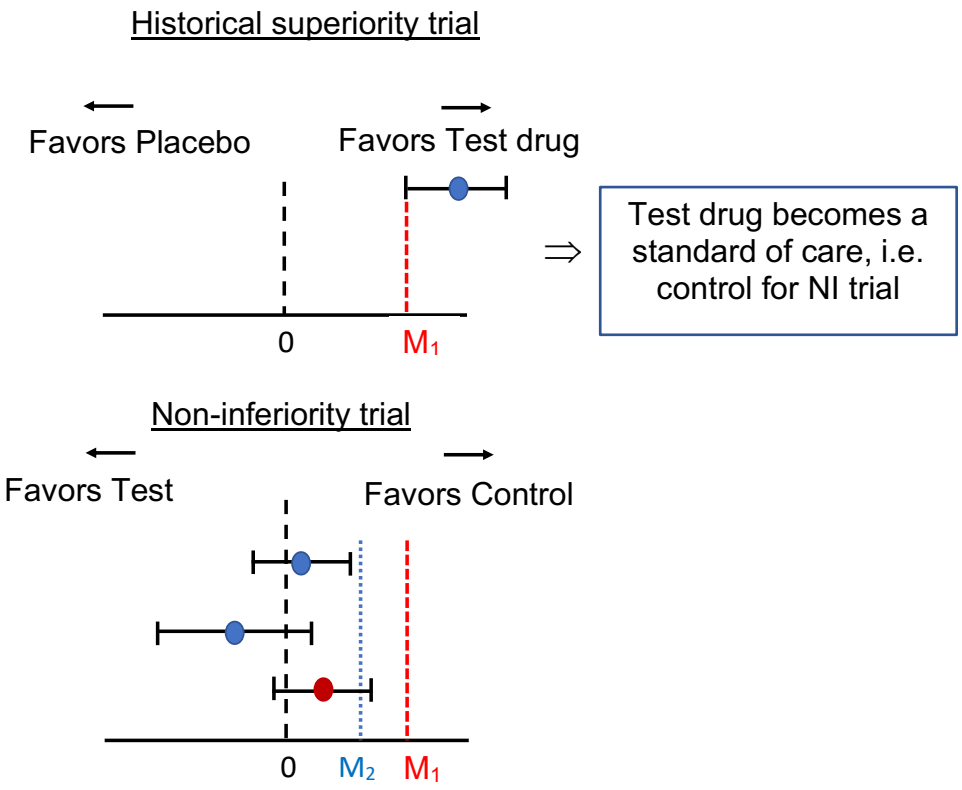

Figure 12: NI clinical trial design with possible outcomes when compared to the standard of care using fixed margin approach. The upper part of the graph presents historical comparison of the standard of care to placebo, while the bottom graph corresponds to the comparison of the new treatment to the standard of care in the non-inferiority trial. NI is concluded for the CIs with blue point estimates, while the CI with a red point estimate corresponds to inferiority of the new test treatment.

As discussed above, M_2 is defined after M_1's determination, which is the previously observed control treatment effect over placebo. Thus, M_2 is commonly seen as a fraction (λ) of the control treatment effect, which clinical experts consider justifiable, i.e., $M_2 = (1 - \lambda)M_1$. We assume that M_1 has been determined based on historical studies and is fixed at the time the non-inferiority trial is being designed, and λ follows some

distribution F with mean μ_λ and variance σ^2_λ. While for a known distribution F, any function of random variable λ can be used to construct the null and alternative hypotheses to test non-inferiority, we will focus on μ_λ throughout this article, since the population mean is a commonly used parameter of interest in many practical situations. Following the notation above we can re-write the hypothesis in (2.1) as:

$$H_0 : p_{Con} - p_{Trt} \geq (1 - \mu_\lambda)M_1 \quad vs \quad H_1 : p_{Con} - p_{Trt} < (1 - \mu_\lambda)M_1 \,. \tag{4.1}$$

For a known population distribution F, we demonstrate how the value of the margin could significantly impact study design in terms of sample size calculation. A sample size per treatment arm (n) can be calculated using the following formula [Blackwelder, 1982, Dann and Koch, 2008, Julious and Owen, 2011], while assuming 1:1 allocation ratio:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2(p_{Con}(1 - p_{Con}) + p_{Trt}(1 - p_{Trt}))}{(p_{Con} - p_{Trt} - (1 - \lambda)M_1)^2} \,, \tag{4.2}$$

where $z_{1-\alpha},\ z_{1-\beta}$ are $1 - \alpha,\ 1 - \beta$ quantiles of standard normal distribution respectively. Specifically, $\alpha,\ 1 - \beta$ represents desired levels of target type-I error and power respectively. Assuming under the alternative hypothesis in (2.1) equality between

the proportions $p_{Con} = p_{Trt}$, given the same type-I error and power, the difference between sample size calculations for some value of $\lambda = \lambda^*$ and μ_λ will be proportional to $\frac{1}{(1-\lambda^*)^2 M_1^2} - \frac{1}{(1-\mu_\lambda)^2 M_1^2}$. This means that, for example, if $p_{Con} = p_{Trt} = 0.8$, $\alpha = 2.5\%$, $1-\beta = 85\%$ and $\mu_\lambda = 0.7$, the sample size per arm using (4.2) for $\lambda = \mu_\lambda$ is 593, while for $\lambda = 0.71$ it would be 634. Thus, a change of just one percent in the amount of the original standard treatment effect to be preserved corresponds in additional 82 subjects to be recruited to an NI study.

The scenario presented here, where the F and its parameters are known is of-course hypothetical and cannot happen in practice. We use it in order to motivate the readers to think about the fraction of the standard treatment effect as of a random variable. Next we discuss how F and it's parameters could be estimated from a survey of clinical experts.

4.2.1 Estimating fraction preservation though a survey

The distribution F and it's parameters μ_λ, σ_λ^2 are considered unknown and ought to be estimated, ideally from a clinical experts survey conducted at the design stage of the trial. We assume that in total K values of λ were collected from clinicians: $\lambda_1, \ldots, \lambda_K$.

Let $\hat{\mu}_\lambda = \frac{1}{K}\sum_{k=1}^{K} \lambda_k$ be MLE of μ_λ. Given a sufficiently large clinical experts survey, the following approximate result holds:

$$\hat{\mu}_\lambda \sim N(\mu_\lambda, \frac{\sigma_\lambda^2}{K}), \tag{4.3}$$

where the variance term can be estimated by $\hat{\sigma}_\lambda^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\lambda_k - \hat{\mu}_\lambda)^2$.

We assume further that the primary analysis of the NI trial is based on Wald's CI as defined in (2.2). Given a sufficiently large sample size per treatment arm, and assuming independence between the NI trial and the clinical experts survey, one can test the hypothesis in (4.1) at α level, by comparing the following upper $(1-\alpha)100\%$ CI with zero:

$$\hat{p}_{Con} - \hat{p}_{Trt} - (1-\hat{\mu}_\lambda)M_1 + z_{1-\alpha}\sqrt{\frac{\hat{p}_{Con}(1-\hat{p}_{Con})}{n_{Con}} + \frac{\hat{p}_{Trt}(1-\hat{p}_{Trt})}{n_{Trt}} + \frac{M_1^2\hat{\sigma}_\lambda^2}{K}}\,. \quad (4.4)$$

If the quantity in (4.4) is smaller than zero, the null hypothesis in (4.1) will be rejected and the new treatment will be declared non-inferior to the standard of care. This approach is, in essence, a synthesis of the information between clinical experts opinions and the data in a new non-inferiority trial. It corresponds to an objective determination of a new treatment's NI, as it takes into account opinions of the multiple clinical experts and the variability associated with such opinions.

The apparent issue with the above approach is that, in practice, it is reasonable to assume that K is small. Therefore the sample of the observed clinical experts responses might not be representative of the clinical experts population, and the normal approximation in (4.3) may not hold.

Although it might be challenging to survey a large number of clinicians to obtain

their opinion about λ, other information related to clinical experts opinions could be more accessible (for example, number of years of treating a disease of interest or number of patients treated), and will be determined as X for the rest of this Chapter. In general, X can be a vector, here for simplicity we will assume that it contains only one random variable. As a result we have a dataset which contains a fully observed X and a partially observed λ. This resembles a missing data problem, which is discussed in the next section.

4.2.2 Treating fraction preservation as missing data

Although observing all the values of λ from a representative experts sample would be extremely helpful and will allow a proper use of (4.3), such observation is unlikely to happen in practice. As a result, we propose to treat unobserved values of λ as missing information. Given additional variable X, which is observed for all the experts from a representative sample, we can use MI procedure to properly estimate μ_λ and σ_λ, which can then be used in (4.4).

For MI purposes, we define a quantity of interest $Q_\lambda = \mu_\lambda$. Similar to the representation in (2.16), we assume that for completely observed values of λ, $(Q_\lambda - \hat{Q}_\lambda) \sim N(0, U_\lambda)$, where $\hat{Q}_\lambda$ is an estimate of Q_λ and U_λ is a variance of $(Q_\lambda - \hat{Q}_\lambda)$. Using a maximum likelihood approach, we have: $\hat{Q}_\lambda = \hat{\mu}_\lambda$ and $U_\lambda = \frac{\hat{\sigma}^2_\lambda}{K}$.

Since it was assumed that demographic and professional characteristics of clinicians affect their opinion about M_2, we used completely observed values of X to multiply

impute the incomplete data. The MI was utilized through classification and regression trees (CART) imputation method [Burgette and Reiter, 2010]. CART is a nonparametric MI approach, where a conditional distribution of a variable is estimated from multiple predictors by forming homogeneous subsets of a predictor space. CART was chosen over a Bayesian linear regression imputation model [Rubin, 2004] due to its tendency to produce small mean squared errors [Akande et al., 2017]. As in Chapter 2, the imputations were produced L times using MICE. The L pairs of estimates $(\hat{Q}_\lambda^{(l)}, U_\lambda^{(l)}),\ (l = 1,\ \ldots, L)$ are then combined following the procedure outlined in (2.17)- (2.21). As a result, we have $(Q_\lambda - \bar{Q}_\lambda)/\sqrt{T_\lambda} \sim t_{\nu\lambda}$, where ν_λ has a similar form as ν in (2.21).

If the subject-level data is fully observed, the $\hat{\mu}_\lambda$ and $\frac{\hat{\sigma}_\lambda^2}{K}$ in (4.4) are then replaced with $\bar{Q}_\lambda$ and T_λ respectively. In addition the $z_{1-\alpha}$ in (4.4) is replaced with an appropriate cut-off value from a sum of normal and Student's t-distribution using general purpose convolution algorithm with Fast Fourier Transformation (FFT) [Kohl et al., 2005, Ruckdeschel et al., 2006].

In case the subject-level data are incomplete, a separate MI procedure should be applied for that data. For simplicity we assume that the incomplete data follow ignorable missingness. Now, we define an additional quantity of interest $Q_Y = p_{Con} - p_{Trt}$, so that for completely observed data $(Q_Y - \hat{Q}_Y) \sim N(0, U_Y)$, where $\hat{Q}_Y = \hat{p}_{Con} - \hat{p}_{Trt}$ and $U_Y = U_{Con} - U_{Trt}$ with $U_i = \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$. Using a logistic regression model with MICE, and observed covariates, the incomplete data is imputed D times. Similar to the margin imputation described above, we will end up with D pairs of estimates $(\hat{Q}_Y^{(d)}, U_Y^{(d)}),\ (d =$

$1, \ldots, D$), which can then be used in the procedure outlined in (2.17)- (2.21). As a result we have: $(Q_Y - \bar{Q}_Y)/\sqrt{T_Y} \sim t_{\nu Y}$, where ν_Y has a similar form as ν_λ above. Following that, in addition to replacing $\hat{\mu}_\lambda$ and $\frac{\hat{\sigma}^2_\lambda}{K}$ with $\bar{Q}_\lambda$ and T_λ in (4.4) respectively, we will also replace the $\hat{p}_{Con} - \hat{p}_{Trt}$ and $\frac{\hat{p}_{Con}(1-\hat{p}_{Con})}{n_{Con}} + \frac{\hat{p}_{Trt}(1-\hat{p}_{Trt})}{n_{Trt}}$ in (4.4) with $\bar{Q}_Y$ and T_Y respectively. Also the $z_{1-\alpha}$ is replaced with an appropriate cut-off value from a sum of two Student's t distribution using the FFT algorithm.

4.2.3 Rates of missing information

Schafer [1997] recommends calculating the rates of missing information, pointing out that such quantities could be useful when evaluating the effect of the incomplete data on the inferential uncertainty of the parameter of interest. In our case, the missingness is due to unobserved clinical experts opinions regarding λ, as well as due to unobserved subject-level data when the patient data are incomplete.

We estimated rates of missing information due to unobserved λ as: $\gamma_\lambda = \frac{B_\lambda}{B_\lambda + \bar{U}_\lambda}$, and rates of missing information due to unobserved subject-level data as $\gamma_Y = \frac{B_Y}{B_Y + \bar{U}_Y}$ [Harel, 2007]. Since, we assume that the two data sources are independent, and the MI is done for each dataset separately, rather than conditionally, the total rate of missing information was defined as $\gamma = \gamma_\lambda + \gamma_Y$.

4.2.4 Simulations details

Subject level information is fully observed

Suppose the overall population of physicians consists out of 1000 medical doctors (MDs), who treat a specific condition. Further, we assume that 300 of these MDs, who are representative of the overall population come to a clinical conference ($K = 300$), and it is feasible for us to survey only 3% of them (9 MDs). Also, we assume that years of experience treating the condition is known for all the MDs, who come to the conference.

Following the above notation, λ_k is a fraction preservation of the control treatment effect over placebo for k^{th} clinical expert, also let X_k be a number of years that clinical expert has been treating a condition of interest. Without loss of generality we will drop the index k from the following explanation. Assume that for any $(\lambda, X) \sim N_2(\mu_\lambda = 0.7, \mu_X = 20, \sigma_\lambda = 0.12, \sigma_X = 7, \rho)$, where $\rho \in (0.4, 0.7)$. The positive correlation between X and λ indicates that more experienced clinical experts are prone to be more conservative with respect to the clinical margin choice. For brevity, and due to similarity between the results, we only present results for $\rho = 0.4$ in this Chapter, while the results for $\rho = 0.7$ appear in Appendix C. Let R_{λ_k} be and indicator variable for whether λ_k was observed ($R_{\lambda_k} = 1$ means that clinician k did not participate in the survey). Two scenarios of participation were considered: i) more experienced clinicians are more likely to participate in the survey; and ii) a random sample from the K clinicians above. For the first scenario, the observed/unobserved values of λ were assigned using $P(R_{\lambda_k} =$

$1|X > 20) = 0.95$ and $P(R_{\lambda_k} = 1|X \leq 20) = 0.99$, while for the second scenario $P(R_{\lambda_k} = 1|X > 20) = P(R_{\lambda_k} = 1|X \leq 20) = 0.97$.

The value of M_1 was set to be 0.23 which was assumed to be known from a meta-analysis of the relevant historical trials. In addition, the subject-level data was generated using a combination of $p_{Con} = 0.8$, $p_{Trt} \in (0.775, 0.8, 0.825)$ and $n_{Con} = n_{Trt} \in (250, 500)$, which resulted in a total of 6 scenarios. The values considered for the simulation are partially based on completed NI trials [Eriksson et al., 2007, 2011]. Each scenario was simulated 5000 times, i.e. both MDs population sample and NI trial data were simulated 5000 times. It should be noted that higher number of simulations did not alter the results.

As stated previously, non-inferiority of the new treatment was determined using the confidence interval in (4.4). The NI decision was considered objective (OBJ) if it was based on the representative sample of MDs (300 MDs). Other methods used for an NI decision were: MI of the margin as described in the previous section with X and $L = 20$, using only observed λ values from the survey (OBS) (only 9 MDs), as well as minimum and maximum values of λ from the representative sample of the K clinicians (MIN and MAX respectively) (one MD each). Minimum and maximum values were considered in order to demonstrate how the NI decision could be affected by consulting only one MD during the conference, who happens to be the least or the most conservative clinician in that conference.

The performance of these methods was assessed by comparing the rates of the NI

decision to the OBJ decision rate. A decision rate was calculated as a proportions of times NI was inferred out of the 5000 simulations. The most favorable approach is the approach for which the NI decision rate is the closet to the OBJ NI decision.

Subject level information is incomplete

After comparing NI decision rates as described in the previous section, where the subject-level information was considered completely observed, we turn to evaluation of NI decision rates when such information is incomplete. For the purposes of this evaluation, we only used survey data where the more experienced MDs were more likely to participate in the survey, a situation that is likely to appear in practice. The incomplete primary outcome data was assumed to follow ignorable missingness, including MCAR and MAR.

In order to impose both MCAR and MAR processes, a variable Z was added to the NI trial simulation. Z was set to have higher values for control treatment group and have higher values for subjects experiencing an event of interest in both groups. Specifically, $Z|Con, Y = 1 \sim N(180, 20)$, $Z|Con, Y = 0 \sim N(100, 20)$, $Z|Trt, Y = 1 \sim N(130, 20)$, $Z|Trt, Y = 0 \sim N(80, 20)$. Z could be seen as a patient reported outcome (PRO) measured during the study, and is positively correlated with the outcome of interest.

Let $R_{S_{ij}}$ be and indicator variable for whether Y_{ij} was observed ($R_{S_{ij}} = 1$ means that outcome Y_{ij} was unobserved for patient j in treatment i). The following logistic function was used to determine observed/unobserved values of Y in each treatment group:

$$P(R_{S_{ij}} = 1) = \frac{1}{1 + exp(-\theta_0 - \theta_{1i} Z_{ij})}, \tag{4.5}$$

where $\theta_0 = log(\frac{DO}{1-DO}) - \theta_{1i}\bar{Z}_i,\ \bar{Z}_i = \sum_{j=1}^{n_i} Z_{ij},\ \theta_{1i}$ represents the effect of Z in group i on the missingness, and DO stands for the overall drop-out rate, which was assumed to be the same in both treatment groups and was set to 20% as a reasonable upper bound for NI trials that encounter some level of missingness [Rabe et al., 2018]. The following two sets of values were considered for θ_{1i}: $\theta_{1,Con} = \theta_{1,Trt} = 0$, which means that the PRO measure Z didn't affect the drop-out of patient j in treatment group i, and $\theta_{1,Con} = -0.009,\ \theta_{1Trt} = 0.013$, which means that patients with lower values in Z were more likely to drop out in the control group, whereas the opposite effect was set in the new treatment group. As a result, the first set of the values for θ_{1i} specified above constituted to MCAR process, while the latter represent an MAR process. Following that, the difference between the two proportion $p_{Con} - p_{Trt}$, was unbiased when estimated from the complete cases under MCAR, and biased under MAR with observed difference being more profound than it actually is.

The incomplete subject-level data was multiply imputed $D = 20$ times as described in Section 4.2.2, and consequently used for NI decision based on MI approach. For OBS/MIN/MAX approaches, the complete cases from the NI trial were used. Similar to Section 4.2.4, the performance of the methods was assessed by comparing the rates

of the NI decision to the OBJ decision rate.

4.3 Results

Figures 13, and 14 demonstrate differences between OBJ NI trial decision and the other four methods under consideration, based on the proportion of the simulated studies that conclude NI when the subject-level data are completely observed. The results in these Figures are presented in terms of the deviation from the OBJ decision. Therefore, the method closest to 0 is considered to be the most favorable. The difference between the Figures is that in Figure 13 MDs who participate in the survey are more likely to be more experienced than MDs who do not participate in the survey, while in Figure 14 that sample is random. As shown in Figures 13-14, the MI approach for NI decision was shown to be the closest to the OBJ decision in most of the scenarios, with deviations between 0.14% and 4.8%. In general, the OBS approach was the second closet to the OBJ, with deviations of between 5.8% and 24%. This was followed by the MIN, which resulted in deviations between 3.4% and 65%. The MAX resulted in the highest deviations, that ranged between 22% and 71%.

Figure 15 has a similar presentation as Figure 13 described above, it demonstrates results for partially observed subject-level data with the MCAR assumption. As shown in Figure 15, the MI-based decision was the closest to the OBJ decision in most of the scenarios, and deviated by 2% to 7.2% from the OBJ rates. In the case where

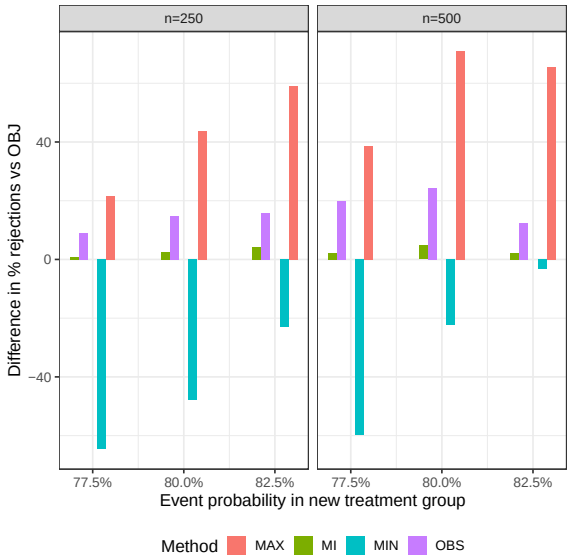

Figure 13: Deviation from objective NI decision, when more experienced MDs are more likely to participate in the survey, subject-level data are fully observed.

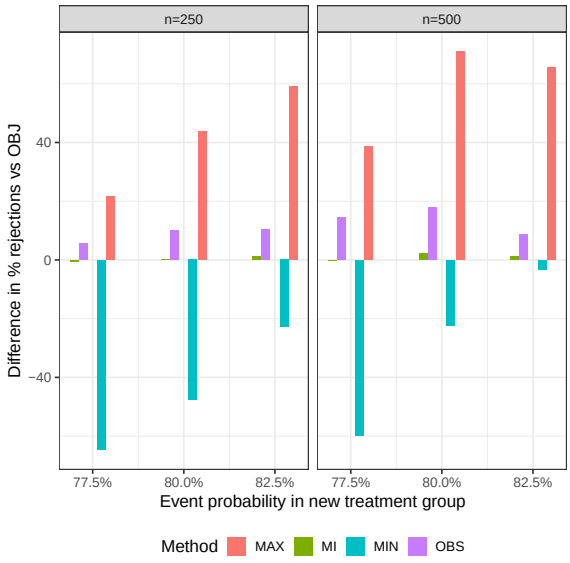

Figure 14: Deviation from objective NI decision, when MDs participation in the survey is completely random, subject-level data are fully observed.

$p_T = 0.825,\ n = 500$, the MIN approach performed similar to MI. To explain this result, we present the NI decision rates (rather than deviations from OBJ) in Table 7. As

show in Table 7, the decision rate for MIN was 100%, which means that all of the 5000 simulated studies concluded NI of the new treatment. This result is not surprising, since, in this case, the new treatment is actually superior by 2.5% to a standard treatment, which means that it would be easier to claim NI. Moreover, the MIN approach represents the least conservative view of the margin, which again would make an NI claim easier to make. For the rest of the scenarios, MIN had over 20% deviation from OBJ decision rates. OBS decision rates deviated between 11% and 31%, while MAX deviation ranged between 22% and 72%.

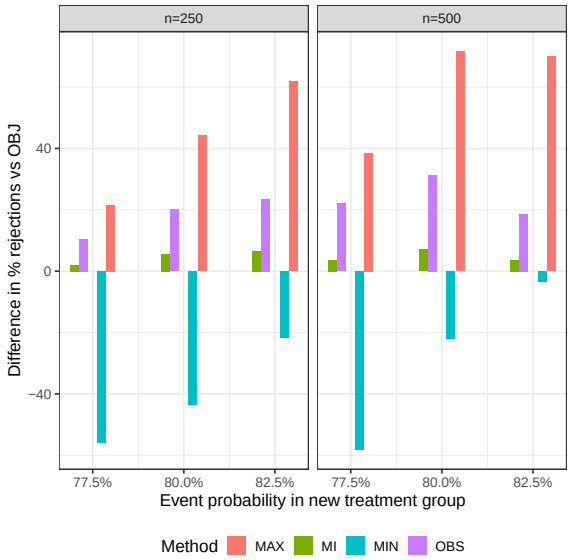

Figure 15: Deviation from population based non-inferiority decision, subject-level data are MCAR.

Figure 16 has a similar presentation as Figure 13 described above, it reveals results for partially observed subject-level data with the MAR assumption. As shown in Figure 16, MI decision approach performed overwhelmingly better than the OBS and the MAX

Table 7: Percent of studies concluding NI by method, when more experienced MDs are more likely to participate in the survey, subject-level data are MCAR.

p_{Trt}	n	OBJ	MI	OBS	MIN	MAX
0.775	250	22.6	20.6	12.1	78.5	1.1
0.775	500	39.4	35.7	17.2	97.5	0.9
0.800	250	49.1	43.5	29.0	92.6	4.8
0.800	500	77.7	70.5	46.4	99.8	6.1
0.825	250	76.9	70.3	53.4	98.7	15.1
0.825	500	96.6	93.0	77.9	100.0	26.5

approaches. Moreover, the deviations from the OBJ decision rates increased dramatically for OBS and MAX. This is reasonable, since the apparent difference in proportions for MAR is larger than it really is, which means that it is harder to claim NI. The MIN approach, however showed similar results to MI for $p_{Trt} = 0.825$ scenarios, as well as $p_{Trt} = 0.8,\ n = 250$.

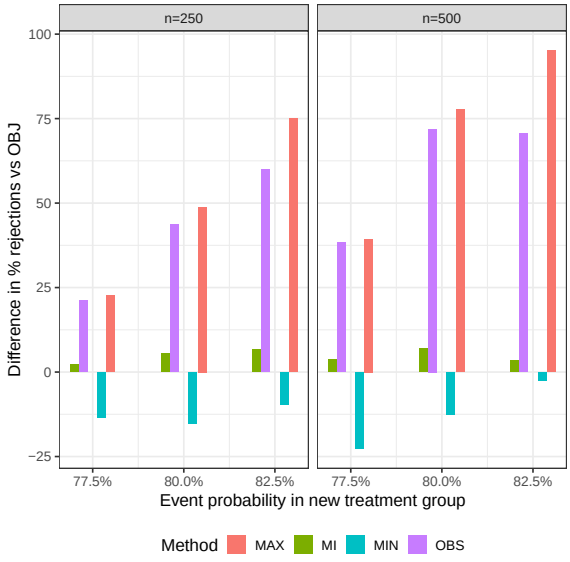

Figure 16: Deviation from population based non-inferiority decision, subject-level data are MAR.

The rates of missing information due to unobserved λ were between 30% and 35%

for both $\rho \in (0.4, 0.7)$ when more experienced MDs were more likely to participate in a survey, and between 27% and 33% when the survey participation was completely random. It should be noted that, as expected, in both cases higher rates of missing information were observed for $\rho = 0.4$. For the incomplete subject-level data, the rates of missing information due to unobserved patient data ranged between 5% and 6% for both MCAR and MAR. As a result, the total rates of missing information were between 35% and 40%. As can be seen, the main contributor to the overall rates of missing information is unobserved clinical experts opinions.

4.4 Conclusion

With NI trial design being more frequently used in recent years, it is imperative to address concerns raised by several systematic reviews of such trials [Wangge et al., 2010, Schiller et al., 2012, Rehal et al., 2016, Althunian et al., 2017, Rabe et al., 2018]. Specifically, one of the major issues that was raised is the apparent lack of justification for the clinically acceptable margin. The choice of such a margin is critical as it directly affects the design stage of an NI study, i.e., the anticipated sample size, as well as further interpretation of the results once the study is complete. Even if, other common issues related to the NI design, such as availability of the historical data and the consistency of standard treatment effect over placebo are resolved, it is still not clear how to choose a clinically acceptable margin.

In this Chapter we present a novel framework, in which we propose to treat the margin as missing information and estimate it through clinical experts survey. Using such framework would allow objective estimation of clinical margin and provide justification for its choice. Furthermore, within this framework we have evaluated the performance of several methods while comparing the rates of NI decisions to the proportions of objective NI determination. Overall, we found that, out of the four approaches considered, decision rates using MI was generally the closest to the objective decision rates. Although, the least conservative margin approach had similar results to MI in several cases, in general, it had high deviations from the OBJ in most of the scenarios. Also, the most conservative choice of clinically acceptable margin deviated significantly from the objective decision rates. Both the most and the least conservative margin choices show the implication and risk of consulting with only one clinical expert, who might have extreme views regarding such a choice. Thus, the determination of the clinically acceptable margin should not be done using an opinion of one clinical expert.

In addition to CART MI, we have also looked at the performance of a more commonly used Bayesian linear regression MI procedure. This method resulted in higher between-imputation variance than CART MI, which is not surprising due to the small number of MDs participating in the survey. As a result, the rate of the simulated studies concluding NI using Bayesian linear regression MI procedure was much further than OBJ decision compared to CART MI. Following that, given the potentially small sample size of the clinical experts survey, we recommend using a CART MI procedure to minimize

subjectivity of the NI decision.

The rates of missing information due to the unobserved clinical experts opinions was the main contributor to the overall rates of missing information. This result underlines the importance of taking into account uncertainty associated with the choice of margin, when it is observed for a small fraction of clinical experts. In addition, it has implications for the study design stage, when the allocation of the study-specific funds is discussed. Following our results, given a limited study budget, an entity running the study might consider allocating a considerable amount of study funds toward the design stage, including margin determination through a clinical experts survey.

We would also like to point out several limitations of this work. First, we only considered a limited number of scenarios. If investigators have a specific scenario in mind which differs from the ones presented here, they should assess it using the framework outline. Second, the framework presented here is new and has not been applied previously, therefore we cannot comment on possible logistic issues that might arise from such data collection besides the ones specified within the framework.

Given the ongoing challenges with respect to NI margin choice and justification, there is a need for a new, more evidence-based, and transparent approach, which takes into considerations variability in clinical experts opinions about such choice. The margin choice has direct implication on the non-inferiority decision, which is important for both drug approval and the public health policy process. However, an approach which allows an objective choice of margin have not been developed prior to this work. Therefore,

the contribution of this Chapter is the novel framework, which accounts for uncertainty associated with non-inferiority margin choice and thus minimizes subjectivity of such choice. Use of this framework will allow an empirical justification of margin choice, and therefore help resolve current practical issues related to it.

Chapter 5

Comprehensive benefit-risk of non-inferior treatments using multi-criteria decision analysis

5.1 Background

As stated in Chapter 1, a non-inferior treatment needs to show some benefit over standard one in order to outweigh decreased effectiveness. As discussed in Chapter 1, several outcomes may be considered when assessing the overall benefit-risk (BR) of the new non-inferior treatment. Various methods exist for structured BR assessment. We chose to use multi-criteria decision analysis (MCDA) due to the reasons mentioned in Chapter 1. One of the MCDA's critiques has been it's deterministic form, which does not take into account variability of either criteria values, or preference weights. To overcome these issues, Tervonen et al. [2011] proposed a stochastic multicriteria acceptability analysis

(SMAA), that also allows BR assessment without explicit weight elicitation from decision makers. Wen et al. [2014] presented two approaches to incorporate clinical data uncertainty into MCDA for BR assessement. Waddingham et al. [2016] presented a Bayesian MCDA model, where outcomes uncertainty was taken into account. Wang et al. [2016] proposed a simpler approach to SMAA, using discriminatory probabilities, which allow comparison between treatments performance for unobserved or partially observed weights. Saint-Hilary et al. [2017] presented an approach that unifies MCDA and SMAA through the use of Direchlet prior.

While the above approaches made a substantial contribution to the structural BR assessment toolkit, these were mainly based on aggregated clinical trial and preference data. In the presence of individual patient-level data, such approaches may be suboptimal since benefits and harm experienced by patients could vary from patient to patient [Evans and Follmann, 2016]. Li et al. [2019] were the first to develop an approach that takes into account heterogeneous responses of the patients to the treatments, as well as consider individual preferences. These authors used a Bayesian multicriteria decision method; an extension of the original SMAA with latent trait models. The main drawback of this approach, however, is its complexity, as well as the lack of clarity how the patient preference weights should be obtained.

The purpose of this work is to develop a simple MCDA approach for structured BR assessment of the new non-inferior treatment when compared to the standard of care using patient-level data from an NI study. With the growing recognition of the importance

of preference elicitation by the patients [Marsh et al., 2017, FDA, 2018], we propose to carry out such elicitation at the beginning of NI trials. Since patient demographic characteristics and baseline (BL) disease status are likely to influence patient outcomes, as well as patient preferences, we believe that it would be beneficial to study these within the same trial. Also, since the introduction of any additional questionnaire is likely to increase burden of the study participants, and investigators, as well as, on the sponsor conducting the study, we suggest gathering preference information only from a random sample of the trial patients, while using MI analysis to create an overall BR assessment.

5.2 Methods

5.2.1 MCDA patient-level indices

Let $MCDA_{ij}$ be MCDA index for patient j ($j = 1, \ldots, n_i$, where n_i is number of patients assigned to treatment i) assigned to treatment i ($i = Con,\ Ttr$). Also let u_{ijc}, and w_{ijc} be a scored outcome/criteria and a preference weight respectively for patient j, for criteria c ($c = 1, \ldots, C$), in treatment group i. We assume that the scored outcomes are calculated using a linear partial value function, which was previously used by several authors [Tervonen et al., 2011, Waddingham et al., 2016, Li et al., 2019]. If ξ_{ijc} is an outcome value for patient j, for outcome c, in treatment i, and ξ'_c, and ξ''_c are the highest and the lowest value to be considered in preference elicitation for outcome c, then $u_{ijc} = \frac{\xi_{ijc} - \xi''_c}{\xi'_c - \xi''_c}$ when higher values are considered as more beneficial, and $u_{ijc} = \frac{\xi'_j - \xi_{ijc}}{\xi'_j - \xi''_j}$

when higher values are considered as more harmful. As a result, all the score values should range between 0 and 1, so that 1 represents the best state and 0 the worst state for a given criteria.

Weight elicitation is assumed to be done using "swing weighting" [Mussen et al., 2007, Marsh et al., 2016]. In a "swing weighting" approach, a subject is required to first determine which criteria's "swing " is the most important to them, and assign 100% to that criteria. "Swing" corresponds to a change from the worst to the best state, for instance, a change from having pain to no pain. Then, the subject needs to provide relative weights to the rest of the considered criteria using "swings" in those criteria. For example, if only treatment response and adverse event (AE) are considered for BR assessment, and a subject identified treatment response as 100%, while AE occurrence as 50%, this means that, for this subject AE occurrence is only half as important as responding to the treatment. It should be noted that in this example, a "swing" in treatment response criteria means experiencing treatment response, while a "swing" in AE criteria means occurrence of AE. After subject-specific weights are obtained, the weights are normalized before they can be used in MCDA calculation.

Following the above, MCDA indices are calculated using:

$$MCDA_{ij} = \sum_{c=1}^{C} u_{ijc}(w_{ijc} / \sum_{c=1}^{C} w_{ijc}) \,. \tag{5.1}$$

An MCDA index for a specific patient represents an overall score for BR trade-off

using the patient's personal preferences. Higher indices contribute to a more favorable BR assessment. Therefore, in order to evaluate overall benefit of a non-inferior treatment, we can compare the MCDA indices between the treatment groups. Since, in drug development, it is not uncommon to compare population means, we chose to use the estimated MCDA means in order to claim whether on average the BR of the new treatment is superior to the available therapy. Due the fact that MCDA indices are constructed using a linear combination of random variables that may have different distributions, it is hard to make a distributional assumption about them. Given the goal of between-group mean comparison, and the central limit theorem (CLT), we chose to make this comparison by constructing 95% CI for mean differences (control vs new treatment) with unequal variances (Welch's T-test). If the upper bound is below zero, then the new treatment would be considered overall more beneficial than the standard treatment. It should be noted that there has been increasing research into the advantages of the simple Welch's T-test over the Wilcoxon-Mann-Whitney test [Skovlund and Fenstad, 2001, Fagerland and Sandvik, 2009, Fagerland, 2012].

5.2.2 MCDA of a random sample of a clinical trial participants

While, ideally, we would like to obtain patient preferences from all trial participants, collecting such additional information might not be feasible as it introduces further burden on patients, investigators and sponsors. We propose to collect patient preferences from only a random sample of the trial participants at the beginning of a trial and use

an MI approach to restore the preferences of all the trial participants. Following that, we assume ignorable missingness throughout this Chapter.

5.2.3 MI for patient preferences

As described above, patient preferences are considered to be obtained using "swinging weighting", which means that the weights could range between 0 and 100. While it is reasonable to require that an imputation procedure produces only values within the plausible range, the ultimate goal of MI is to efficiently estimate a parameter of interest. Therefore, it is not necessary to multiply impute values within a specified range. In fact, Rodwell et al. [2014] performed a simulation study, where different MI approaches were implemented for bounded incomplete variables. Specifically the authors evaluated: the commonly used Bayesian linear regression with no truncation, post-imputation processing (i.e. assign min/max if the imputed values falls outside the plausible range), truncated regression and predictive mean matching. The authors concluded that, in terms of bias of the point estimate, variance and coverage, the usual regression with no truncation performed better than the other methods.

Given a variety of MI methods [Harel and Zhou, 2007], and results from Rodwell et al. [2014] we decided to use Bayesian linear regression (NORM). Due to its simplicity, we also evaluated Bayesian linear regression with post-processing (NORM TRUNCATED). In addition, we assessed CART MI [Burgette and Reiter, 2010]. We chose to use CART due to its ability to capture complex data structures [Burgette and Reiter, 2010], which

might arise from patient preference elicitation. In addition, it should be noted, that CART returns imputed values which are within the range of the observed values, therefore the weights range of 0 to 100 is satisfied.

Similar to the representation in (2.16), we assume that $(Q - \hat{Q} \sim N(0, U)$, where Q is the mean difference in MCDA indices between the standard and new non-inferior treatment. Specifically, in our case $\hat{Q} = \overline{MCDA}_{Con} - \overline{MCDA}_{Trt}$, where $\overline{MCDA}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} MCDA_{ij}$. Also, $U = \frac{S^2_{MCDA_{Con}}}{n_{Con}} + \frac{S^2_{MCDA_{Trt}}}{n_{Trt}}$, where $S^2_{MCDA_i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(MCDA_{ij} - \overline{MCDA}_i)^2$. Further, we assume that patient preferences are imputed L times, and within each l^{th} imputed dataset $(l = 1, \ldots, L)$ we obtain pairs of summaries defined above $(\hat{Q}^{(l)}, U^{(l)})$. Using Rubin's combination rules Rubin [2004] as described in Chapter 2, the L pairs of estimates are then combined following the procedure outlined in (2.17)-(2.21). As a result, we have $(Q - \bar{Q})/\sqrt{T} \sim t_\nu$, where ν has a similar form as ν in (2.21).

5.2.4 Simulation

Simulation set-up for clinical trial data

Simulation set-up is based on the information from recently published results of NI HAWK study Dugel et al. [2019]. HAWK was one of the two NI pivotal studies, that showed the NI of Brolucizumab 6mg when compared to Aflibercept 2mg for treatment of neovascular age-related macular degeneration. The planned sample size for this study was 297 patients per treatment arm, therefore we used 300 patients per arm for simplicity

in our simulations.

We simulated the following efficacy outcomes for primary and a key secondary endpoint: best corrected visual acuity (BCVA) change from BL to week 48, and central subfield thickness (CST) change from BL to week 16. It should be noted that, while higher values of BCVA correspond to a better health, the opposite is true for CST. In addition, we simulated the following safety variables, which were presented by Dugel et al. [2019]: non-ocular adverse events (AEs), and ocular AEs.

According to Dugel et al. [2019], the analysis of primary endpoint was adjusted for age and BCVA at BL, thus the primary endpoint (PE) outcome was generated as a linear function of these two BL measures. The statistical analysis methods for the key secondary endpoints (SEs) were not specified by the authors. Therefore, similar to the primary outcome, the SE was generated as a linear function of age and CST value at BL. For simulation purposes, it was assumed that all the outcome measures are more likely to worsen for elder patients, as well as unfavorable events are more likely to appear in elder patients. In addition, for efficacy variables, we assumed that patients who are sicker at BL would have lower rates of disease worsening than patients who are healthier at BL. Also, we assumed that patients with lower BCVA are more likely to experience ocular AEs.

The simulated BL values included: BCVA, CST, subretinal fluid (SRF), intraretinal fluid (IRF), sub presence of retinal pigment epithelium (sub-RPE), age and gender. SRF, IRF and sub-RPE BL variables were added, because these were used in other key

SEs, and thus should be considered as important. The BL variables were generated using a multivariate normal distribution, with a vector of mean and standard deviation values as reported in [Dugel et al., 2019], for binary variables the reported proportion of events (p) was used as mean and $\sqrt{p(1-p)/n}$, where n is a sample size was used as standard deviation. Since the vision-related measures deteriorate with age, positive pairwise correlations were assumed between age, and CST, SRF, IRF, sub-RPE, whereas negative pairwise correlation was assumed between age and BCVA. Also, negative pairwise correlations between BCVA and CST/SRF/IRF/sub-RPE were assumed, while positive pairwise correlations between CST and SRF/IRF/sub-RPE were set. For simplicity it was assumed, that SRF, IRF and sub-RPE had zero pairwise correlations, and gender was assumed uncorrelated with other variables.

Simulation set-up for patient weights

To assess the overall benefit of the new non-inferior treatment to a standard treatment, preference weights were assigned using patients BL characteristics. Also continuous BL characteristics (BCVA and CST) were categorized using cut-off values reported in [Dugel et al., 2019]. The weights were generated using the following three scenarios:

1. In Scenario 1, we used three criteria: BCVA change from BL to week 48 (primary endpoint (PE)), ocular and non-ocular AEs. It was assumed that, on average, patients care more about non-ocular AEs than PE or ocular AEs. Preference weights for PE were defined as a function of BCVA at BL as follows: mean weight

of (70, 50, 30) were assigned to patients with lower, medium and high values of BCVA at BL respectively. This indicates that a patient with worse vision status at BL will be more likely to assign higher weights for improvement of the vision status, than one with a better vision at BL. For ocular AEs, the mean weights of (50, 80) were assigned for female and male patients respectively. For non-ocular AEs, the mean weights of (70, 90) were assigned for female and male patients respectively. Higher weights for AEs for male patients indicate that men care more about experiencing AEs than women.

2. In Scenario 2, we used four criteria: PE, ocular AEs, non-ocular AEs, and CST change from BL to week 16. In this scenario, patients on average care more about PE, than the other three endpoints. The mean weights for PE were assigned as (90, 60, 30) for low, medium and high values of BCVA at BL respectively. The mean weights for ocular AEs, and non-ocular AEs were set to (50, 70) for female and male patients respectively. Also, the weights for CST endpoint were defined mean values of (30, 45) for patients with low and high CST at BL respectively. This indicates that patients with low CST care less about changes in this outcome.

3. In Scenario 3, we used the same criteria as in Scenario 2 above. The weights for PE were defined in the same way as in Scenario 2. For ocular AEs, the mean weights of (70, 80) were assigned for female and male patients respectively, while the mean weights of (30, 40) for females and male patients were assigned respectively for

non-ocular AEs. For CST, the mean weights of (15, 30) were assigned for patients with low and high CST at BL respectively. In this scenario, patients, on average, care more about PE, and ocular AEs than non-ocular AEs and changes in CST.

We compared MCDA indices for all the enrolled patients, then we randomly masked between 50% to 90% of the patients preference, and consequently MCDA indices (employing MCAR missingness structure) overall and made the same comparison using CCA or MI. The idea here was to see whether MI would be as good as using all study participants, so that performing preference elicitation as an ancillary study within an NI study is advantageous. In addition, for Scenario 3, we have also evaluated masking between 50% to 90% of patients preferences based on patients' CST scores at BL (employing MAR missingness structure) so that patients with low CST at BL have higher probability of being unobserved than patients with high CST at BL.

Patient weights were imputed using all available BL characteristics. It should be noted that BCVA and CST at BL were used in their continuous form. Similar to the previous chapters, all the MI methods were implemented using MICE. The number of simulations was 1000, the number of imputations was 10, and CART MI was set for a minimum of 10 leaves per split. It should be noted that increasing the number of simulations or imputations, as well as decreasing the number of leaves per split, did not alter the inferences.

5.3 Results

Results of a single simulated study are presented in Figure 17. Each column of the graph corresponds to a different criteria/outcome, with the top part of the column showing descriptive statistics of the criteria at the end of the trial by treatment group, and the bottom part of the column showing the distribution of the weights associated with that criteria and assigned by the patients at BL. In addition, different colors on the bottom part of each column correspond to the BL characteristics that affect patients' preferences. For example, the first column in Figure 17 corresponds to BCVA change from BL criteria (which is the PE), the top part of the graph shows boxplot for PE by treatment. As can be seen there is no difference between the treatments in terms of BCVA, which is in accordance with the results reported in [Dugel et al., 2019]. The bottom part of the graph shows the distribution of the weights assigned for the PE by the patients based on BCVA at BL status (status can be low, medium or high, which mean severe, moderate and mild vision impairment respectively). As specified above, we assume that patients with worse vision at BL are more likely to assign higher weights for vision improvement. And this is what is seen in the bottom part of the graph, i.e., patients with "low" (red) BCVA at BL have higher weight values, while patients with "med" (green) and "high" BCVA at BL assign lower weights. The second and third columns of Figure 17 correspond to similar representation of ocular and non-ocular AEs, whereas the preferences for both are affected by gender as specified in the previous Section. Overall, Figure 17 shows that

the new treatment seems to be better than the control in non-ocular AEs, this criteria also has higher weights than the other two. Following that, the new treatment might have only a slightly better BR profile than the control overall.

Figure 18 presents results from the 1000 simulated studies, and percent of the time the new non-inferior treatment was declared as favorable when compared to control using the three outcomes as specified in Scenario 1 and their corresponding weights assigned by the patients. The dashed line in the Figure represents a result when the weights are obtained for all the study participants, while the columns show results produced by the methods under evaluation when only using a sample of the trial participants. As can be seen in Figure 18, the MI results are close to the fully observed weights, while CCA results are not. For example, when only 10% of the trial participants provide their preferences, then CCA will declare a new non-inferior treatment as favorable only 5.1% of the time, while CART MI, NORM MI, and NORM TRUNCATED MI will do so 17.4%, 17.9%, and 16.2% of the times respectively. Given the fact that 17.6% correspond to the fully observed weights (dashed line), the MI methods demonstrate favorable results.

Figure 19 has a similar presentation as Figure 17, and shows results of a single study for Scenario 2. As shown in Figure 19, the patients care more about PE than the other three outcomes. However given favorable results in CST outcome and non-ocular AEs, there is a good chance that the new treatment has a better BR profile than a control. This is demonstrated in Figure 20, which has a similar presentation as Figure 18, here the % of the studies with a favorable BR for the new treatment is 93.1%. Also, the MI

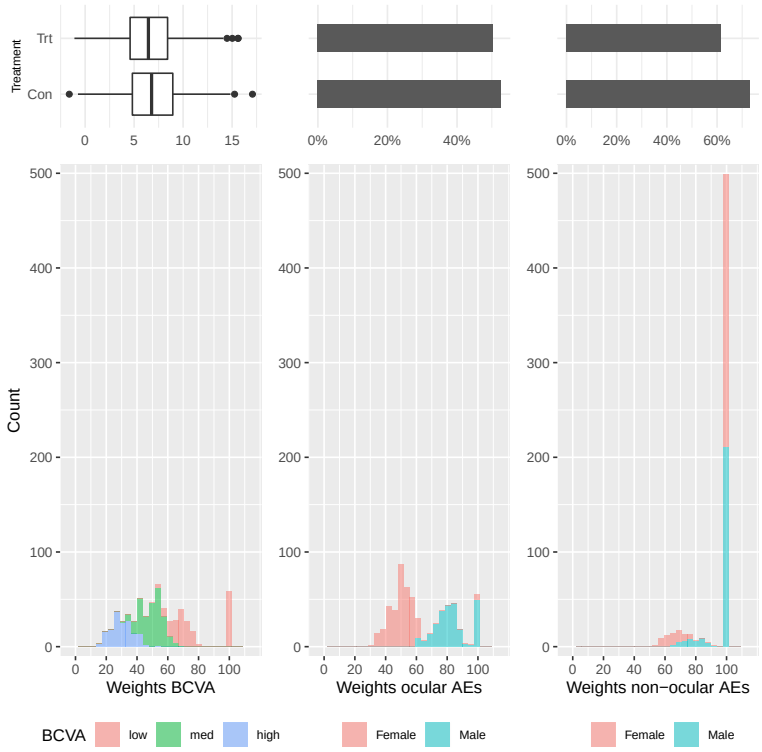

Figure 17: Outcome criteria values per treatment group (upper plots) and weights distribution (lower plots). Single study results. Scenario 1.

results are close to the fully observed weights, which is not the case for CCA. There is also a minimal difference between the MI methods.

Figure 21 has a similar presentation as Figure 17, and shows results of a single study for Scenario 3. As shown in Figure 21, the preferences for the change in CST, the outcome that shows a clear advantage of the new treatment over the control, are lower than in Scenario 2 (Figure 19). As a result, it is expected that the overall BR profile is now worse than in Scenario 2 (Figure 20), which is shown in Figure 22. The % of the studies with a favorable BR of the new treatment is now 32.1%. The results of MI and CCA are consistent with the previous two scenarios.

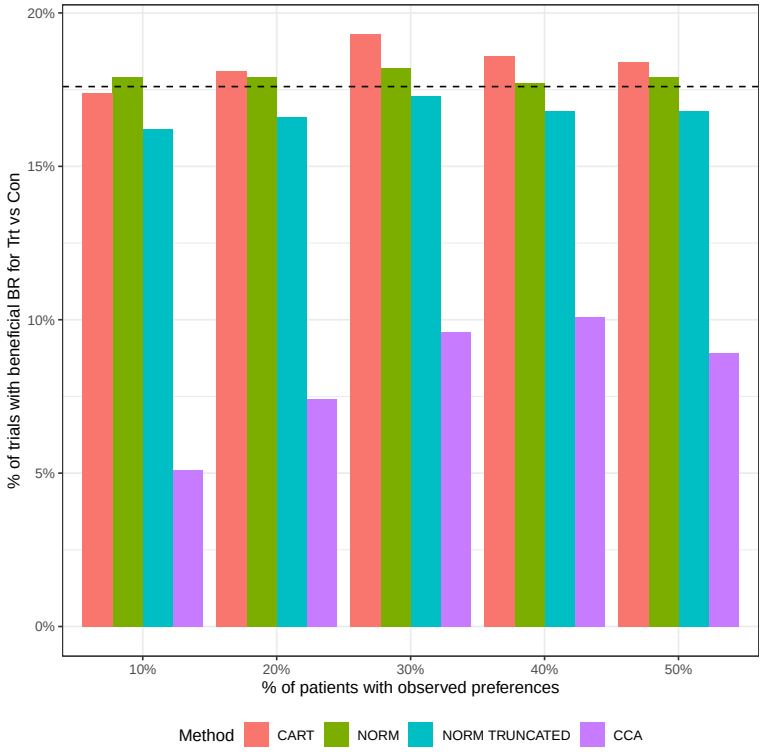

Figure 18: % of trials with beneficial BR profile for a new treatment. Scenario 1. The dashed line represents a result when the preference weights are observed for all study participants.

Figure 23 has similar presentation as Figure 18, and shows percent of the times the new non-inferior treatment is declared as beneficial compared to the control, when MAR structure is imposed in Scenario 3. As can be seen in Figure 23, the results from MI were still favorable, with minimal differences between the imputation methods.

5.4 Discussion

In this Chapter, we present a novel approach for comprehensive BR assessment of new non-inferior treatments which allows researchers to elicit patient preferences from an NI

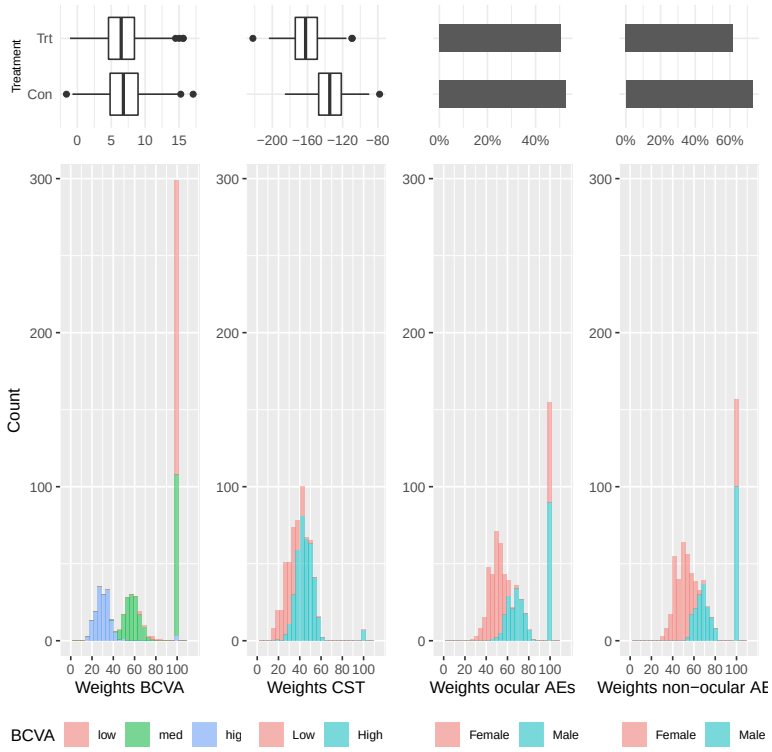

Figure 19: Outcome criteria values per treatment group (upper plots) and weights distribution (lower plots). Single study results. Scenario 2.

study. Similar to the original MCDA, our approach allows one to incorporate multiple outcomes into BR assessment. The advantage of our approach over the original MCDA, is that it also allows one to incorporate variability associated with both outcomes values and patient preferences. In order to minimize a participation burden for patients and medical monitors, as well as a financial burden for a sponsor due the collection of the preference data, we propose collecting such data only for a sample of trial participants. If such a sample is random, then as we showed, MI results are very close to the results based on all trial participants preferences.

While our simulations were mainly based on MCAR, we also showed favorable results

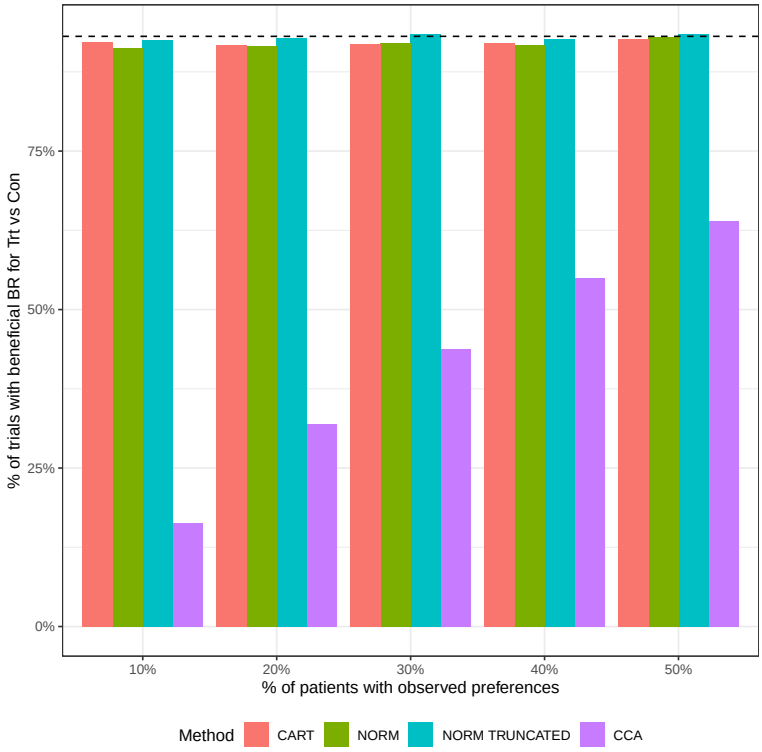

Figure 20: % of trials with beneficial BR profile for a new treatment. Scenario 2. The dashed line represents a result when the preference weights are observed for all study participants.

of MI under MAR. It should be noted, that although MCAR is unlikely to hold in clinical trials [Little et al., 2012], this missingness structure could be embedded in a study design. In other words, one could design a study where the participants of the ancillary preference survey are randomly chosen. If the preferences are influenced by patients' BL characteristics, a plausibility of MCAR can be evaluated by comparing these characteristics between ancillary study participants and non-participants.

The consideration of MCDA scores' variability in BR assessment irrespective of a trial design has been recently evaluated by Wen et al. [2014] and by Broekhuizen et al. [2017].

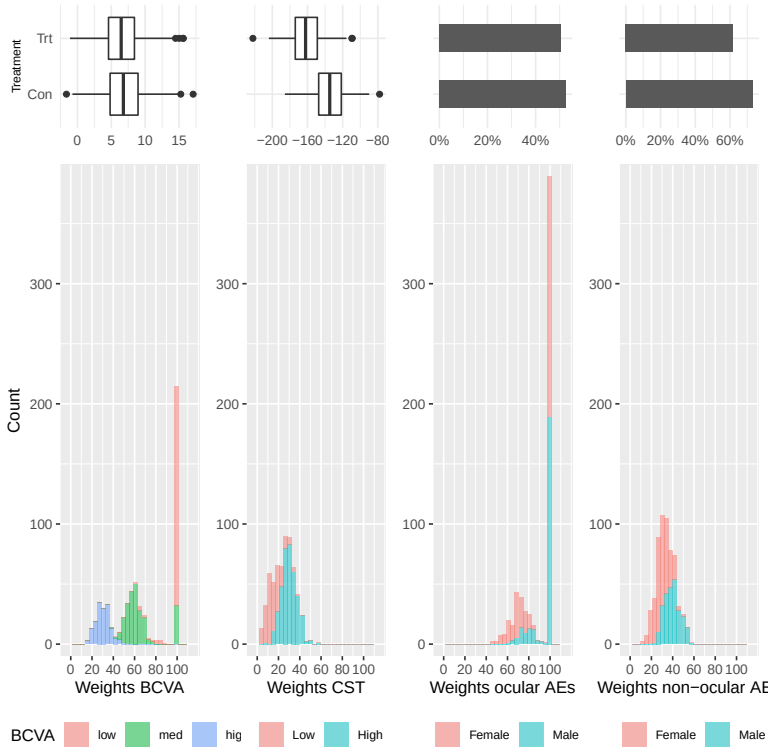

Figure 21: Outcome criteria values per treatment group (upper plots) and weights distribution (lower plots). Single study results. Scenario 3.

Wen et al. [2014] proposed two methods that take into account clinical data variability, including a simulation method. Although the authors focused on the aggregated clinical trial data, they outlined a bootstrap procedure which could be used in cases when patient-level data are available. It should be noted that patients preferences in this approach are fixed. We implemented the bootstrap procedure as proposed by Wen et al. [2014], while using fixed weights in Scenario 3 from Section 5.2. The fixed weights were calculated as weighted average of the mean preferences specified for Scenario 3 based on the relevant BL characteristics. For example, since the mean weights for the PE were (90, 60, 30) for low, medium and high values of BCVA at BL respectively, and given for

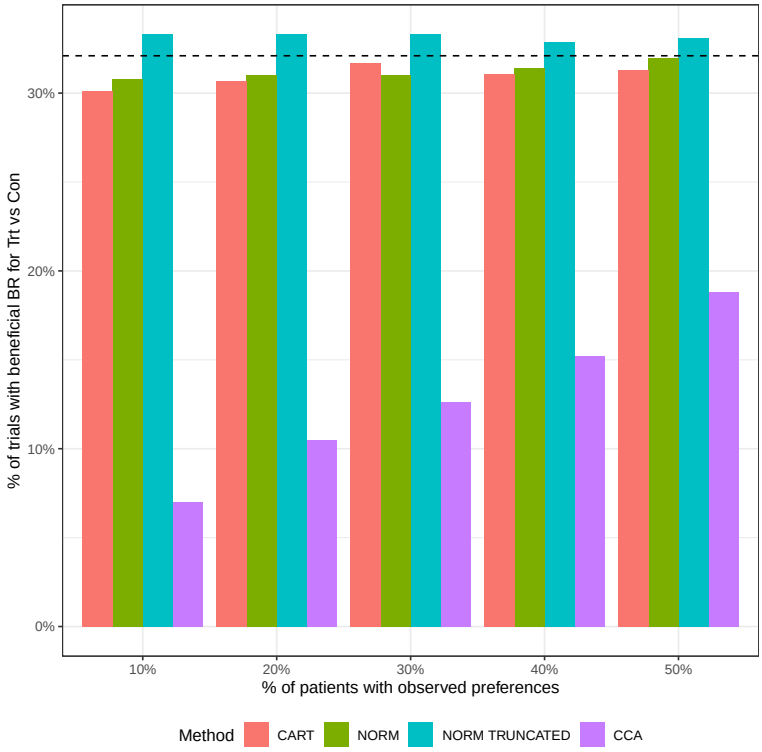

Figure 22: % of trials with beneficial BR profile for a new treatment. Scenario 3. The dashed line represents a result when the preference weights are observed for all study participants.

example, proportion of patients who have such BCVA values at BL is 20%, 50% and 30% respectively, the fixed weight for that PE for all the patients would be 84. The use of one fixed weight for all the patients, rather then individual patient preferences resulted in 52.9% of the simulated studies concluding favorable BR of a non-inferior treatment over a standard one. This is 20% more optimistic than the original 32.1% (dashed line in Figure 22), when each patient provided his/her preference.

Broekhuizen et al. [2017] suggested a probabilistic approach, where uncertainty in weight preferences identified by the patients in the previous preference survey studies,

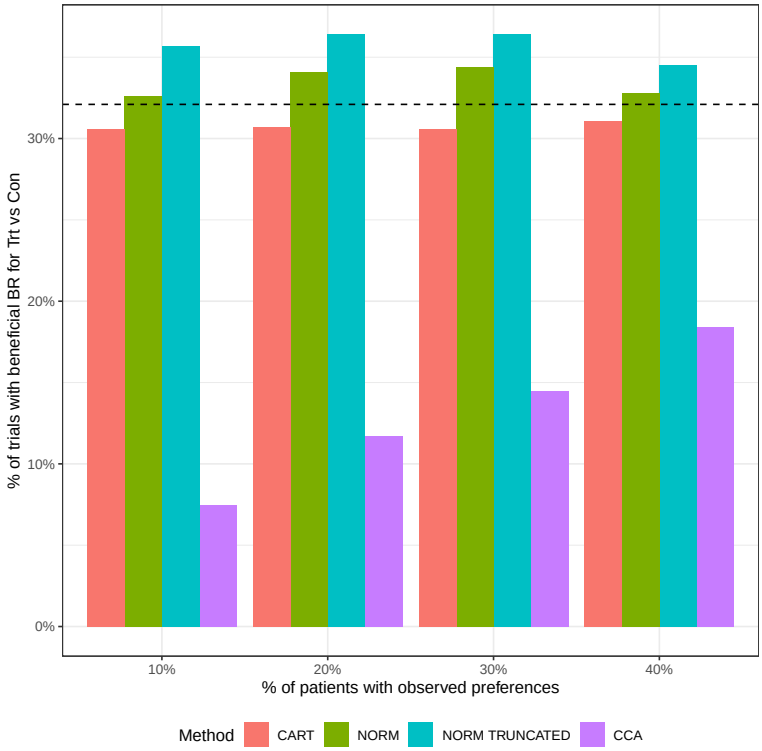

Figure 23: % of trials with beneficial BR profile for a new treatment. Scenario 3 with MAR missingness structure. The dashed line represents a result when the preference weights are observed for all study participants.

as well as uncertainty in the clinical data are taken into account using Markov Chain Monte Carlo simulation. We believe that this approach could be useful if the preference survey results are available prior to new NI trial initiation, and if both survey and NI trial participants come from the same patient population. In order to evaluate a possible implication of the differences between survey and NI trial participants, we simulated Scenario 3 from Section 5.2 assuming that survey participants are sicker than those in NI trial. Specifically, we assumed that 70%, 20% and 10% of the survey participants have low, medium and high BCVA BL values respectively, as well as 90% have high CST

values. Since we anticipate patients' preferences to be affected by BL characteristics, encountering sicker patients will translate into higher BCVA and CST preferences. Using the mean values of the new preferences along with the outcome values of the study participants in accordance with Broekhuizen et al. [2017], we received 42.8% of simulated studies concluding favorable BR of the non-inferior treatment. This is 10% more optimistic than the original 32.1%.

The above two comparisons underline the importance of taking into consideration both outcome and preferences variability, as well as obtaining preferences from the same patient population as being investigated during an NI study. While BR assessment is critical for consideration of any new therapy, it is especially important for NI trials. Given an acceptable clinical worsening in effectiveness of non-inferior treatment, there is a need for formal assessment of the treatment's advantages over a standard of care. If the non-inferior treatment does not offer benefits that outweigh decreased effectiveness, there is no justification for its approval. However, so far such formal BR assessment was not used for non-inferior treatments. Hence, our novel approach provides a pragmatic solution for BR assessment of non-inferior treatments.

The limitations of our study include the simulations that are limited to the scenarios being considered in this Chapter. In addition, we used a partial linear function for scoring of outcome criteria. The use of other function types is outside the scope of this work.

The contribution of this Chapter is a development of a new tool for comprehensive

BR assessment of non-inferior treatments. Since our approach is based on the data collected during the study and therefore reported in the clinical study report, it would facilitate a more transparent BR evaluation in practice. Moreover, incorporation of the individual patients' perspectives from the target patient population is aligned with recent regulatory commitments to include patient preferences in BR assessment. Our method contributes to a better decision making process with regards to new non-inferior treatments.

Chapter 6

Conclusion

This dissertation focuses on new approaches for the design and analysis of NI clinical trials. Each Chapter presented in this dissertation looks at different statistical issues related to this type of trial. Specifically, in Chapter 2, we provide a set of recommendations for incomplete data analysis along with a novel approach for analysis of data under MNAR. This contributes towards better analysis practices of NI trials. In Chapter 3, we introduce MI combination rules for difference in binomial proportions, when NW method is used. While we use this method for analysis of NI trials and therefore contribute again towards better analysis of such trials, it is a general methodology which is useful for other applications as well. In Chapter 4 we present a new framework for incorporating different clinical experts opinions regarding NI margin into tthe design and analysis of NI trials. While in Chapter 5, we develop a simple BR assessment approach for evaluation of overall benefit of non-inferior treatment. Both Chapter 4 and 5 advance design, analysis and interpretation of NI clinical trials. This dissertation provides an important contribution to the field of Statistics, and drug development. The

novel methods and techniques outlined in this dissertation facilitate practitioners involved with NI trails to make more efficient and transparent evaluations of treatment effectiveness.

Appendix A

A.1 Sample size per scenario and method

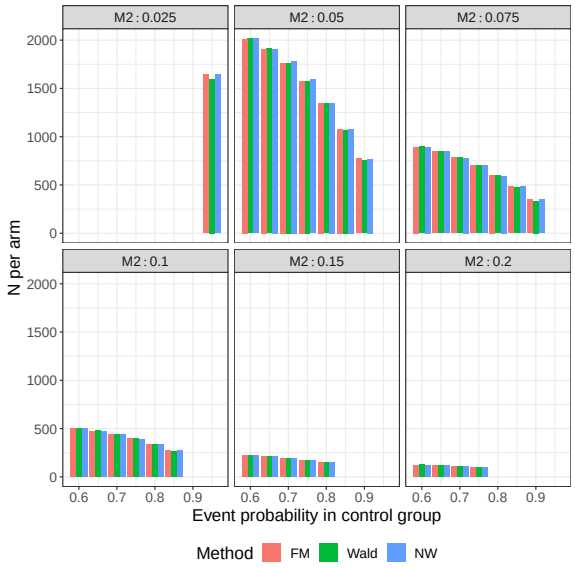

Figure A1: Sample size per scenario and method

A.2 Outcome variable model

The baseline covariates were set to follow bivariate Normal distribution: $(X_1, X_2) \sim N_2(\mu_1 = 4, \mu_2 = 100, \sigma_1 = 1, \sigma_2 = 20, \rho = -0.3)$, where μ_1, μ_2 represent mean values for X_1, X_2 respectively, σ_1, σ_2 represent standard deviation values for X_1, X_2 respectively, and ρ is a correlation coefficient between X_1 and X_2. X_1 could for example represent

disease status at baseline, while X_2 could be a systolic blood pressure at baseline.

Model parameters in (8) were set as following: $\beta_{Grpout} = log\frac{p_{Trt}*(1-p_{Con})}{p_{Con}*(1-p_{Trt})}$, $\alpha_y = log\frac{p_{Con}}{(1-p_{Con})} - \beta_{1out} * \mu_1 - \beta_{2out} * \mu_2$, and $\beta_{1out} = 0.1, \beta_{2out} = -0.01$. The values of β_1, β_2 were found through simulation and were calibrated to achieve the target proportions of favorable events in the treatment arms.

A.3 Additional results for Chapter 2

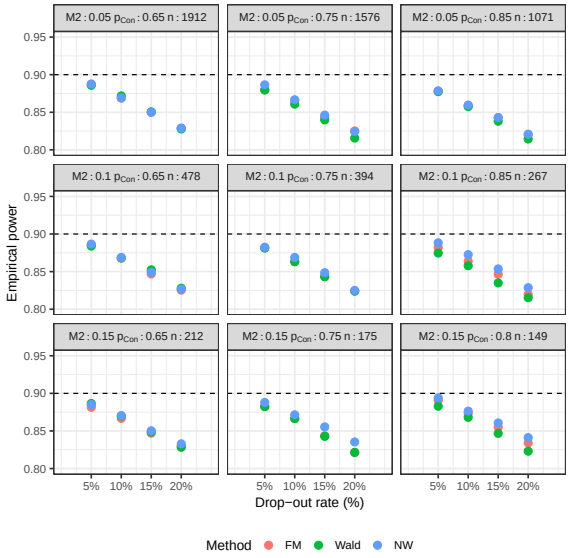

Figure A2: Empirical power CCA imputation strategy for MCAR: drop-out rates are balanced between the treatment groups

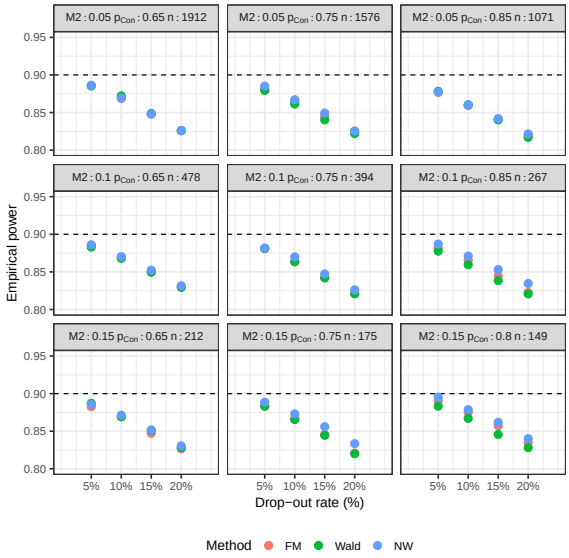

Figure A3: Empirical power CCA strategy for MAR: drop-out rates are balanced between the treatment groups

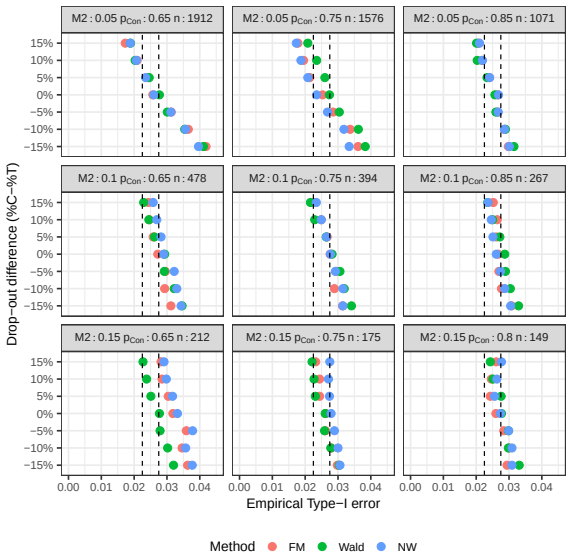

Figure A4: Empirical type-I error CCA strategy for MAR, overall drop-out rate of 20%

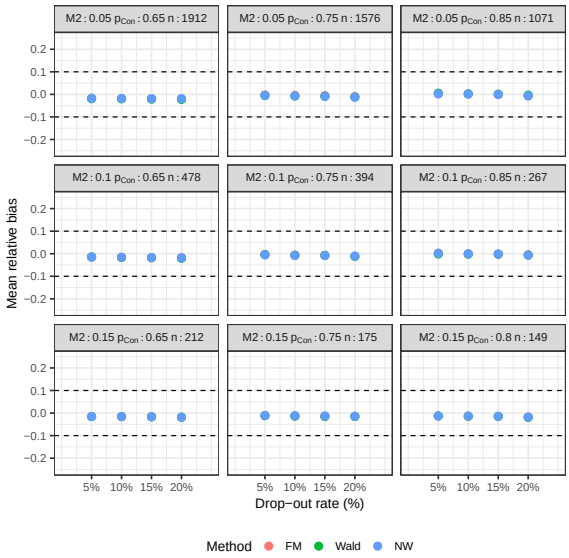

Figure A5: Mean relative bias CCA strategy for MAR: drop-out rates are balanced between the treatment groups

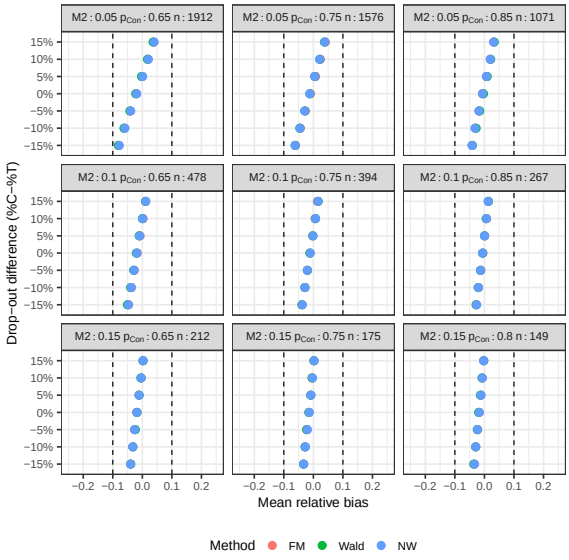

Figure A6: Mean relative bias CCA strategy for MAR, overall drop-out rate of 20%

Table A1: Mean relative bias for MNAR due to lack of efficacy in Trt for different drop-out (DO) rates, CCA and two-stage MI strategies, Wald method

p_{Con}	DO	M_2	CCA	MI
0.65	0.05	15%	-0.697	-0.050
0.65	0.10	15%	-0.351	-0.002
0.65	0.15	15%	-0.230	0.008
0.75	0.05	15%	-0.669	-0.056
0.75	0.10	15%	-0.357	-0.014
0.75	0.15	15%	-0.248	0.031
0.80	0.15	15%	-0.257	0.002
0.85	0.05	15%	-0.566	-0.082
0.85	0.10	15%	-0.332	-0.011
0.65	0.05	10%	-0.483	-0.052
0.65	0.10	10%	-0.244	-0.013
0.65	0.15	10%	-0.158	-0.001
0.75	0.05	10%	-0.468	-0.058
0.75	0.10	10%	-0.247	-0.019
0.75	0.15	10%	-0.174	0.015
0.80	0.15	10%	-0.181	-0.007
0.85	0.05	10%	-0.399	-0.073
0.85	0.10	10%	-0.233	-0.019
0.65	0.05	5%	-0.258	-0.041
0.65	0.10	5%	-0.130	-0.014
0.65	0.15	5%	-0.088	-0.009
0.75	0.05	5%	-0.248	-0.042
0.75	0.10	5%	-0.129	-0.015
0.75	0.15	5%	-0.095	-0.001
0.80	0.15	5%	-0.099	-0.011
0.85	0.05	5%	-0.210	-0.046
0.85	0.10	5%	-0.122	-0.014

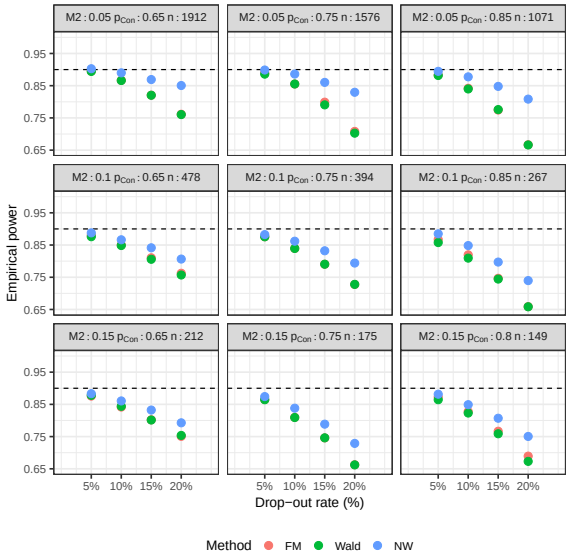

Figure A7: Empirical power two-stage MI strategy for MNAR due to lack of efficacy in Trt

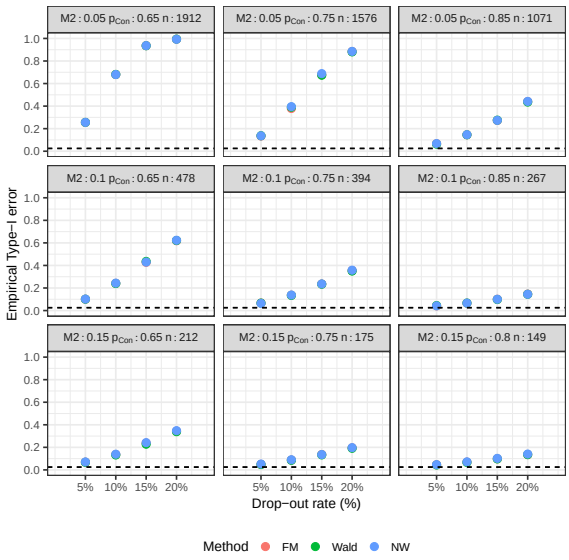

Figure A8: Empirical type-I errors, CCA strategy for MNAR due to overwhelming efficacy in Con

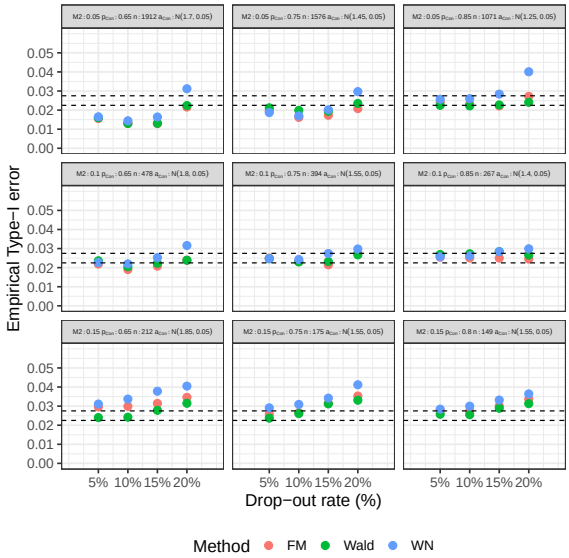

Figure A9: Empirical type-I errors, two-stage MI strategy via MICE for MNAR due to overwhelming efficacy in *Con*

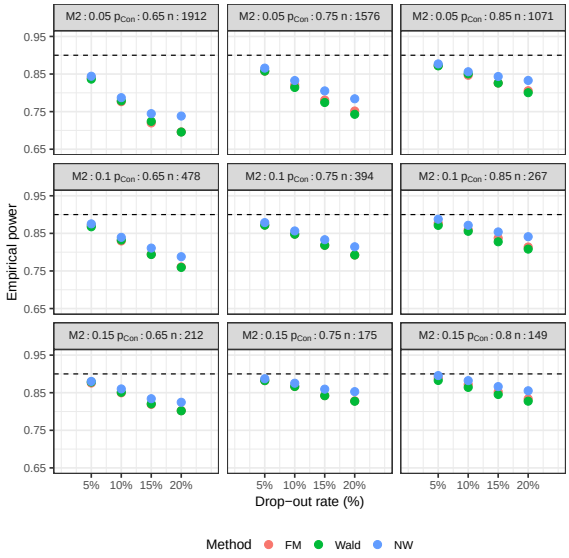

Figure A10: Empirical power two-stage MI strategy for MNAR due to overwhelming efficacy in *Con*

Table A2: Mean relative bias for MNAR due to overwhelming efficacy in Con for different drop-out (DO) rates, CCA and two-stage MI strategies, Wald method

p_{Con}	M_2	DO	CCA	MI
0.65	0.05	20%	-1.650	0.134
0.65	0.10	20%	-0.831	0.093
0.65	0.15	20%	-0.560	0.043
0.75	0.05	20%	-1.185	0.098
0.75	0.10	20%	-0.594	0.076
0.75	0.15	20%	-0.406	0.017
0.80	0.15	20%	-0.330	0.028
0.85	0.05	20%	-0.689	0.050
0.85	0.10	20%	-0.347	0.063
0.65	0.05	15%	-1.235	0.147
0.65	0.10	15%	-0.624	0.083
0.65	0.15	15%	-0.421	0.040
0.75	0.05	15%	-0.871	0.097
0.75	0.10	15%	-0.437	0.065
0.75	0.15	15%	-0.301	0.016
0.80	0.15	15%	-0.244	0.020
0.85	0.05	15%	-0.497	0.051
0.85	0.10	15%	-0.250	0.047
0.65	0.05	10%	-0.825	0.118
0.65	0.10	10%	-0.418	0.060
0.65	0.15	10%	-0.284	0.028
0.75	0.05	10%	-0.569	0.077
0.75	0.10	10%	-0.287	0.044
0.75	0.15	10%	-0.201	0.010
0.80	0.15	10%	-0.161	0.012
0.85	0.05	10%	-0.318	0.041
0.85	0.10	10%	-0.161	0.033
0.65	0.05	5%	-0.419	0.059
0.65	0.10	5%	-0.214	0.026
0.65	0.15	5%	-0.149	0.009
0.75	0.05	5%	-0.280	0.041
0.75	0.10	5%	-0.142	0.021
0.75	0.15	5%	-0.104	0.001
0.80	0.15	5%	-0.084	0.001
0.85	0.05	5%	-0.150	0.026
0.85	0.10	5%	-0.077	0.017

Appendix B

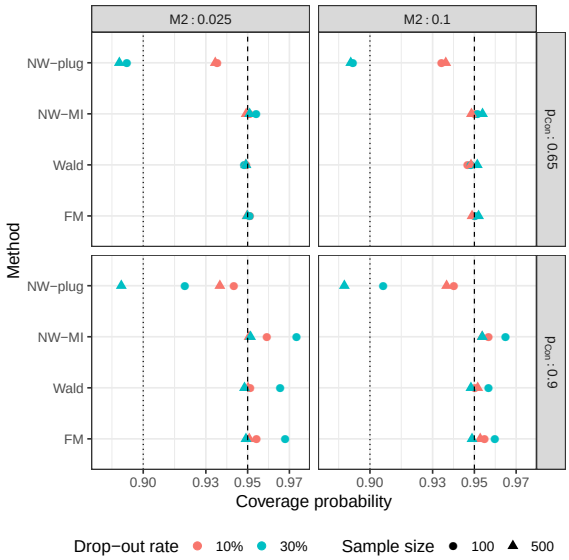

Figure B1: Coverage probability for MAR with independent X and Y (Dashed line represents the desired coverage probability of .95, dotted line represents covarage probability of .90.)

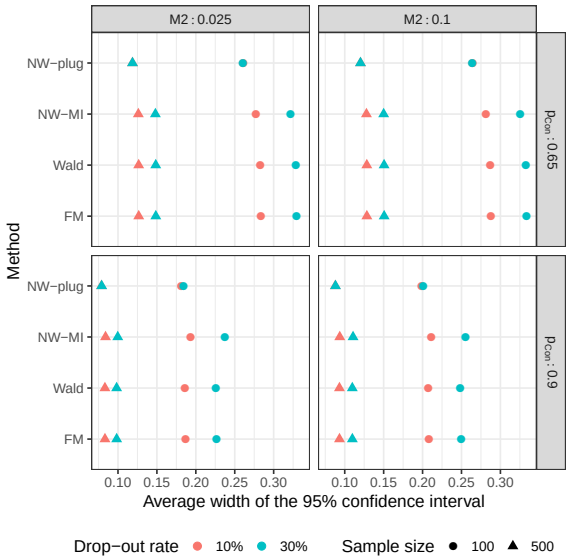

Figure B2: Average width of 95% confidence intervals for MAR, with independent X and Y

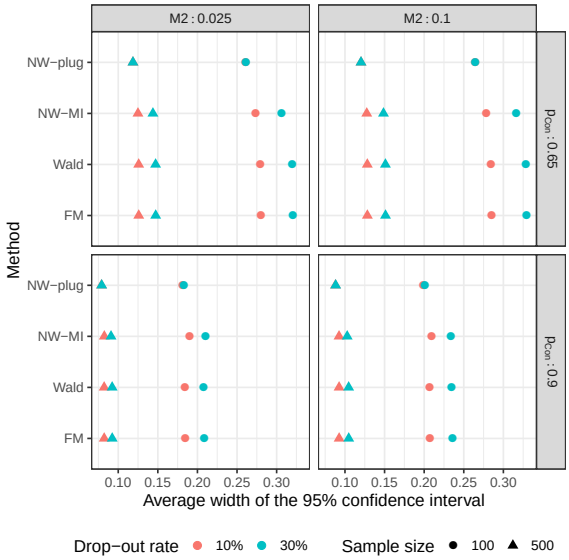

Figure B3: Average width of 95% confidence intervals for MNAR

Appendix C

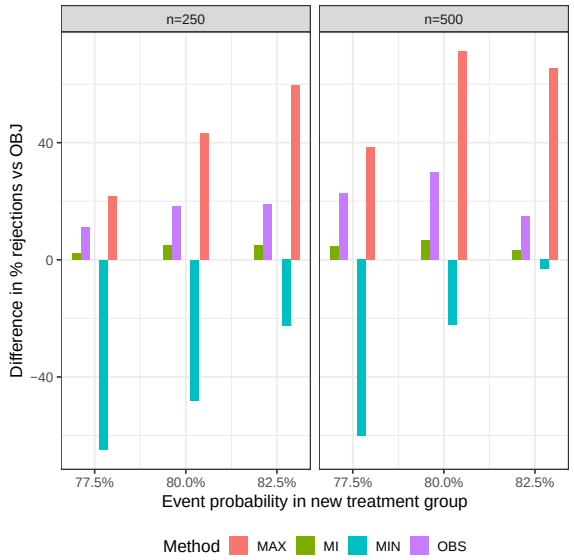

Figure C1: Deviation from objective NI decision, when more experienced MDs are more likely to participate in the survey, subject-level data are fully observed, $\rho = 0.7$.

Table C1: Percent of studies concluding NI by method, when more experienced MDs are more likely to participate in the survey, subject-level data are MCAR, $\rho = 0.7$.

p_{Trt}	n	OBJ	MI	OBS	MIN	MAX
0.775	250	22.8	19.4	10.3	78.7	1.1
0.775	500	39.1	32.9	15.0	97.4	0.9
0.800	250	48.8	41.2	25.6	92.4	4.7
0.800	500	77.9	68.7	42.0	99.8	6.2
0.825	250	77.2	69.1	49.5	98.7	15.1
0.825	500	96.9	92.7	74.7	100.0	26.6

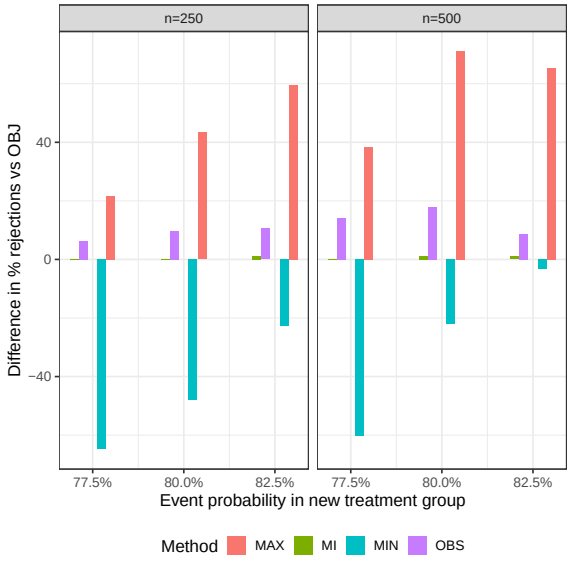

Figure C2: Deviation from objective NI decision, when MDs participation in the survey is completely random, subject-level data are fully observed, $\rho = 0.7$.

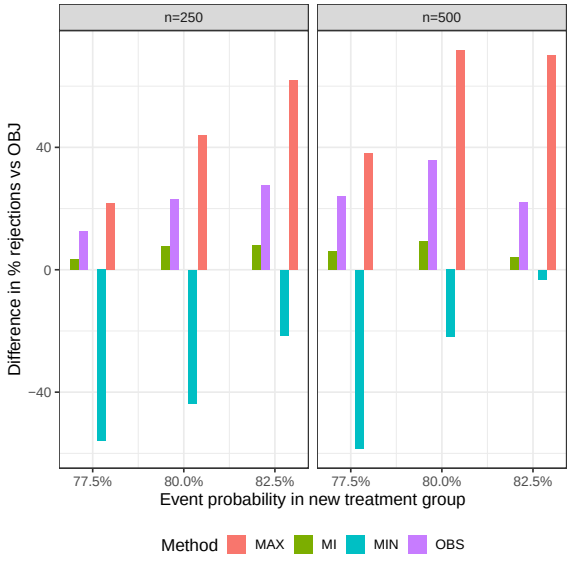

Figure C3: Deviation from population based non-inferiority decision, subject-level data are MCAR, $\rho = 0.7$.

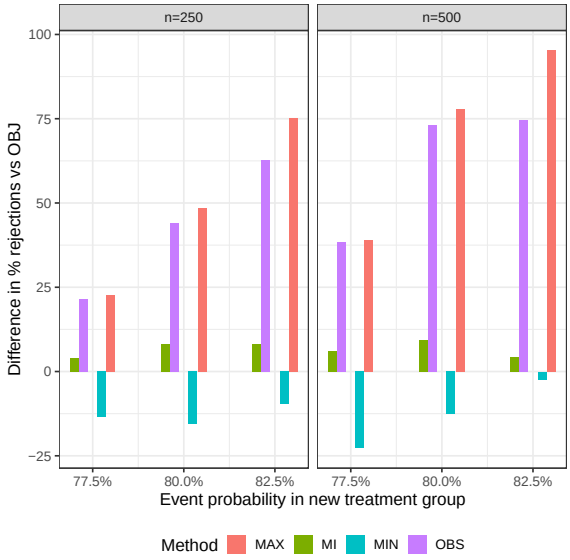

Figure C4: Deviation from population based non-inferiority decision, subject-level data are MAR, $\rho = 0.7$.

Bibliography

Scott K Aberegg, Andrew M Hersh, and Matthew H Samore. Empirical consequences of current recommendations for the design and interpretation of noninferiority trials. *Journal of general internal medicine*, pages 1–9, 2017.

Alan Agresti and Brian Caffo. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4):280–288, 2000.

Olanrewaju Akande, Fan Li, and Jerome Reiter. An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71 (2):162–170, 2017.

Félix Almendra-Arao. A study of the classical asymptotic noninferiority test for two binomial proportions. *Drug Information Journal*, 43(5):567–571, 2009.

Turki A Althunian, Anthonius de Boer, Olaf H Klungel, Widya N Insani, and Rolf HH Groenwold. Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. *Trials*, 18(1):107, 2017.

Takeshi Amemiya. Tobit models: A survey. *Journal of econometrics*, 24(1-2):3–61, 1984.

GA Barnard. A new test for 2× 2 tables. *Nature*, 156:177, 1945.

Jonathan W Bartlett, Ofer Harel, and James R Carpenter. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology*, 182(8):730–736, 2015.

William C Blackwelder. "proving the null hypothesis" in clinical trials. *Controlled clinical trials*, 3(4):345–353, 1982.

David R Bristol. Superior safety in noninferiority trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(1):75–81, 2005.

Henk Broekhuizen, Maarten J IJzerman, A Brett Hauber, and Catharina GM Groothuis-Oudshoorn. Weighing clinical evidence using patient preferences: an application of probabilistic multi-criteria decision analysis. *PharmacoEconomics*, 35(3):259–269, 2017.

Lawrence Brown and Xuefeng Li. Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference*, 130(1-2):359–375, 2005.

Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076, 2010.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

An-Wen Chan, Jennifer M Tetzlaff, Peter C Gøtzsche, Douglas G Altman, Howard Mann, Jesse A Berlin, Kay Dickersin, Asbjørn Hróbjartsson, Kenneth F Schulz, Wendy R Parulekar, et al. Spirit 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj*, 346:e7586, 2013.

Ivan SF Chan. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in medicine*, 17(12):1403–1413, 1998.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

CMPH. Guideline on missing data in confirmatory clinical trials. *London: European Medicines Agency*, 2010.

Linda M Collins, Joseph L Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.

Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC, 2008.

Rebekkah S Dann and Gary G Koch. Methods for one-sided testing of the difference between proportions and sample size considerations related to non-inferiority clinical trials. *Pharmaceutical statistics*, 7(2):130–141, 2008.

Hakan Demirtas and Joseph L Schafer. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in medicine*, 22(16): 2553–2575, 2003.

Pravin U Dugel, Adrian Koh, Yuichiro Ogura, Glenn J Jaffe, Ursula Schmidt-Erfurth, David M Brown, Andre V Gomes, James Warburton, Andreas Weichselberger, Frank G Holz, et al. Hawk and harrier: phase 3, multicenter, randomized, double-masked trials of brolucizumab for neovascular age-related macular degeneration. *Ophthalmology*, 2019.

James D Dziura, Lori A Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, 86(3):343, 2013.

Hans-Georg Eichler, Eric Abadie, June M Raine, and Tomas Salmonson. Safe drugs and the cost of good intentions. *New England Journal of Medicine*, 360(14): 1378–1380, 2009.

EMEA. Benefit-risk methodology project. *London: European Medicines Agency*, 2009.

EMEA. Work package 1 report: Description of the current practice of benefit-risk assessment for centralised procedure products in the eu regulatory network. 2011.

Bengt I Eriksson, Ola E Dahl, Nadia Rosencher, Andreas A Kurth, C Niek van Dijk, Simon P Frostick, Martin H Prins, Rohan Hettiarachchi, Stefan Hantel, Janet Schnee, et al. Dabigatran etexilate versus enoxaparin for prevention of venous thromboembolism after total hip replacement: a randomised, double-blind, non-inferiority trial. *The Lancet*, 370(9591):949–956, 2007.

Bengt I Eriksson, Ola E Dahl, Michael H Huo, Andreas A Kurth, Stefan Hantel, Karin Hermansson, Janet M Schnee, Richard J Friedman, RE-NOVATE II Study Group, et al. Oral dabigatran versus enoxaparin for thromboprophylaxis after primary total hip arthroplasty (re-novate ii). *Thrombosis and haemostasis*, 105 (04):721–729, 2011.

Scott R Evans and Dean Follmann. Using outcomes to analyze patients rather than patients to analyze outcomes: a step toward pragmatism in benefit: risk evaluation. *Statistics in biopharmaceutical research*, 8(4):386–393, 2016.

Morten W Fagerland. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC Medical Research Methodology*, 12(1):78, 2012.

Morten W Fagerland and Leiv Sandvik. The wilcoxon–mann–whitney test under scrutiny. *Statistics in medicine*, 28(10):1487–1497, 2009.

Conor P Farrington and Godfrey Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine*, 9(12):1447–1454, 1990.

FDA. Structured approach to benefit-risk assessment in drug regulatory decision-making. *Draft PDUFA V implementation plan—February 2013: fiscal years 2013–2017*, 2014.

FDA. Non-inferiority clinical trials to establish effectiveness; guidance for industry, 2016.

FDA. Benefit-risk assessment in drug regulatory decision making. *Draft PDUFA VI Implementation Plan. FY2018–2022*, 2018.

Ronald Aylmer Fisher. The design of experiments. 1935.

Thomas R Fleming. Current issues in non-inferiority trials. *Statistics in medicine*, 27(3):317–332, 2008.

Thomas R Fleming. Addressing missing data in clinical trials. *Annals of internal medicine*, 154(2):113–117, 2011.

Paul Gallo and Christy Chuang-Steiny. A note on missing data in noninferiority trials. *Drug Information Journal*, 43(4):469–474, 2009.

Silvio Garattini, Vittorio Bertele, and Luca LiBassi. Placebo or active control? either, as long as it is in the patient's interest. *WHO Drug Information*, 17(4): 253, 2003.

Beryl Primrose Gladstone and Werner Vach. Analyzing noninferiority trials: it is time for advantage deficit assessment–an observational study of published noninferiority trials. *Open Access Journal of Clinical Trials*, 7:11, 2015.

John W Graham, Bonnie J Taylor, Allison E Olchowski, and Patricio E Cumsille. Planned missing data designs in psychological research. *Psychological methods*, 11 (4):323, 2006.

Jeff J Guo, Swapnil Pandey, John Doyle, Boyang Bian, Yvonne Lis, and Dennis W Raisch. A review of quantitative risk–benefit methodologies for assessing drug safety and efficacy—report of the ispor risk–benefit management working group. *Value in Health*, 13(5):657–666, 2010.

Ofer Harel. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4(1):75–89, 2007.

Ofer Harel and Xiao-Hua Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

HM James Hung and Sue-Jane Wang. Statistical considerations for noninferiority trial designs without placebo. *Statistics in Biopharmaceutical Research*, 5(3):239–247, 2013.

HM James Hung, Sue-Jane Wang, Yi Tsong, John Lawrence, and Robert T O'Neil. Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine*, 22(2):213–225, 2003.

HM James Hung, Sue-Jane Wang, and Robert O'Neill. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(1):28–36, 2005.

HM James Hung, Sue-Jane Wang, and Robert O'Neill. Issues with statistical risks for testing methods in noninferiority trial without a placebo arm. *Journal of Biopharmaceutical Statistics*, 17(2):201–213, 2007.

HM James Hung, Sue-Jane Wang, and Robert O'Neill. Challenges and regulatory experiences with non-inferiority trial design without placebo arm. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):324–334, 2009.

ICH. International conference on harmonisation. statistical principles for clinical trials e9, 1998.

ICH. International conference on harmonisation. choice of control group and related issues in clinical trials e10, 2000.

ICH. Estimands and sensitivity analysis in clinical trials, 2017.

Steven A Julious and Roger J Owen. A comparison of methods for sample size estimation for non-inferiority studies with binary outcomes. *Statistical methods in medical research*, 20(6):595–612, 2011.

Matthias Kohl, Peter Ruckdeschel, and Thomas Stabla. General purpose convolution algorithm for distributions in s4-classes by means of fft. Technical report, Citeseer, 2005.

Kan Li, Sheng Luo, Sammy Yuan, and Shahrul Mt-Isa. A bayesian approach for individual-level drug benefit-risk assessment. *Statistics in medicine*, 2019.

Zhengqing Li and Christy Chuang-Stein. A note on comparing two binomial proportions in confirmatory noninferiority trials. *Drug information journal*, 40(2): 203–208, 2006.

Ilya Lipkovich and Brian L Wiens. The role of multiple imputation in non-inferiority trials for binary outcomes. *Statistics in Biopharmaceutical Research*, (just-accepted), 2017.

Roderick J Little, Ralph D'agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.

Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

Todd D Little and Mijke Rhemtulla. Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7(4):199–204, 2013.

Qing Liu, Yulan Li, and Katherine Odem-Davis. On robustness of noninferiority clinical trial designs against bias, variability, and nonconstancy. *Journal of biopharmaceutical statistics*, 25(1):206–225, 2015.

Anne Lott and Jerome P Reiter. Wilson confidence intervals for binomial proportions with multiple imputation for missing data. *The American Statistician*, pages 1–7, 2018.

Kevin Marsh, Maarten IJzerman, Praveen Thokala, Rob Baltussen, Meindert Boysen, Zoltán Kaló, Thomas Lönngren, Filip Mussen, Stuart Peacock, John Watkins, et al. Multiple criteria decision analysis for health care decision making—emerging good practices: report 2 of the ispor mcda emerging good practices task force. *Value in health*, 19(2):125–137, 2016.

Kevin Marsh, J Jaime Caro, Alaa Hamed, and Erica Zaiser. Amplifying each patient's voice: a systematic review of multi-criteria decision analyses involving patients. *Applied health economics and health policy*, 15(2):155–162, 2017.

Olli Miettinen and Markku Nurminen. Comparative analysis of two rates. *Statistics in medicine*, 4(2):213–226, 1985.

Shahrul Mt-Isa, Nan Wang, Christine E Hallgreen, Torbjörn Callréus, Georgy Genov, Ian Hirsch, Steve Hobbiger, Kimberley S Hockley, Davide Luciani, Lawrence D Phillips, et al. Review of methodologies for benefit and risk assessment of medication. *London, UK: PROTECT Consortium*, 2014.

Shahrul Mt-Isa, Mario Ouwens, Veronique Robert, Martin Gebel, Alexander Schacht, and Ian Hirsch. Structured benefit–risk assessment: a review of key publications and initiatives on frameworks and methodologies. *Pharmaceutical statistics*, 15(4):324–332, 2016.

Venkatesh L Murthy, Nihar R Desai, Amit Vora, and Deepak L Bhatt. Increasing proportion of clinical trials using noninferiority end points. *Clinical cardiology*, 35 (9):522–523, 2012.

Filip Mussen, Sam Salek, and Stuart Walker. A quantitative approach to benefit-risk assessment of medicines–part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiology and drug safety*, 16(S1):S2–S15, 2007.

Robert G Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in medicine*, 17(8):873–890, 1998.

Tie-Hua Ng. Noninferiority hypotheses and choice of noninferiority margin. *Statistics in medicine*, 27(26):5392–5406, 2008.

Masako Nishikawa, Toshiro Tango, and Megu Ohtaki. Statistical tests based on new composite hypotheses in clinical trials reflecting the relative clinical importance of multiple endpoints quantitatively. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(5):749–762, 2009.

NRC. *The prevention and treatment of missing data in clinical trials*. National Academies Press, 2011.

Egon S Pearson. The choice of statistical tests illustrated on the interpretation of data classed in a 2× 2 table. *Biometrika*, 34(1/2):139–167, 1947.

Lawrence D Phillips et al. Benefit-risk methodology project: work package 2 report: applicability of current tools and processes for regulatory benefit-risk assessment. 2011a.

Lawrence D Phillips et al. Benefit-risk methodology project: work package 3 report: field tests. 2011b.

Lawrence D Phillips et al. Benefit-risk methodology project: work package 4 report: benefit-risk tools and processes. 2012.

Lawrence D Phillips et al. Benefit-risk methodology project: update on work package 5: effects table pilot (phase i). 2014.

Gilda Piaggio, Diana R Elbourne, Stuart J Pocock, Stephen JW Evans, Douglas G Altman, Consort Group, et al. Reporting of noninferiority and equivalence randomized trials: extension of the consort 2010 statement. *Jama*, 308(24):2594–2604, 2012.

Milo A Puhan, Sonal Singh, Carlos O Weiss, Ravi Varadhan, and Cynthia M Boyd. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC medical research methodology*, 12(1):173, 2012.

Brooke A Rabe, Simon Day, Mallorie H Fiero, and Melanie L Bell. Missing data handling in non-inferiority and equivalence trials: A systematic review. *Pharmaceutical Statistics*, 2018.

Russell Reeve. Confidence interval of difference of proportions in logistic regression in presence of covariates. *Statistical methods in medical research*, 27(2):451–465, 2018.

Sunita Rehal, Tim P Morris, Katherine Fielding, James R Carpenter, and Patrick PJ Phillips. Non-inferiority trials: are they inferior? a systematic review of reporting in major medical journals. *BMJ open*, 6(10):e012594, 2016.

Jerome P Reiter and Trivellore E Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480): 1462–1471, 2007.

Laura Rodwell, Katherine J Lee, Helena Romaniuk, and John B Carlin. Comparison of methods for imputing limited-range variables: a simulation study. *BMC medical research methodology*, 14(1):57, 2014.

Peter Roebruck and Andreas Kühn. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine*, 14(14):1583–1594, 1995.

Joachim Röhmel, Christoph Gerlinger, Norbert Benda, and Jürgen Läuter. On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(6):916–933, 2006.

Mark D Rothmann, Brian L Wiens, and Ivan SF Chan. *Design and analysis of non-inferiority trials*. Chapman and Hall/CRC, 2016.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Peter Ruckdeschel, Matthias Kohl, Thomas Stabla, and Florian Camphausen. S4 classes for distributions. r news 6 (2): 2–6, 2006.

Gaelle Saint-Hilary, Stephanie Cadour, Veronique Robert, and Mauro Gasparini. A simple way to unify multicriteria decision analysis (mcda) and stochastic multicriteria acceptability analysis (smaa) using a dirichlet distribution in benefit–risk assessment. *Biometrical Journal*, 59(3):567–578, 2017.

Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

Petra Schiller, Nicole Burchardi, Michael Niestroj, and Meinhard Kieser. Quality of reporting of clinical non-inferiority and equivalence randomised trials-update and extension. *Trials*, 13(1):214, 2012.

Zijin Shen. *Nested Multiple Imputations*. PhD thesis, Harvard University, 2000.

Juned Siddique, Ofer , and Catherine M Crespi. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *The annals of applied statistics*, 6(4):1814, 2012.

Juned Siddique, Ofer Harel, Catherine M Crespi, and Donald Hedeker. Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty: application to a smoking cessation trial. *Statistics in medicine*, 33 (17):3013–3028, 2014.

Yulia Sidi and Ofer Harel. The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209:169–173, 2018.

Eva Skovlund and Grete U Fenstad. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? *Journal of clinical epidemiology*, 54(1):86–92, 2001.

Katie J Suda, Anne M Hurley, Trevor McKibbin, and Susannah E Motl Moroney. Publication of noninferiority clinical trials: changes over a 20-year interval. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 31(9): 833–839, 2011.

Tommi Tervonen, Gert van Valkenhoef, Erik Buskens, Hans L Hillege, and Douwe Postmus. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428, 2011.

Praveen Thokala, Nancy Devlin, Kevin Marsh, Rob Baltussen, Meindert Boysen, Zoltan Kalo, Thomas Longrenn, Filip Mussen, Stuart Peacock, John Watkins, et al. Multiple criteria decision analysis for health care decision making—an introduction: report 1 of the ispor mcda emerging good practices task force. *Value in health*, 19(1):1–13, 2016.

TORPA. The organisation for professionals in regulatory affairs. *Regulatory Rapporteur*, 9(6), 2012.

Eline van Overbeeke, Chiara Whichello, Rosanne Janssens, Jorien Veldwijk, Irina Cleemput, Steven Simoens, Juhaeri Juhaeri, Bennett Levitan, Jürgen Kübler, Esther de Bekker-Grob, et al. Factors and situations influencing the value of patient preference studies along the medical product lifecycle: a literature review. *Drug discovery today*, 24(1):57–68, 2019.

Ed Waddingham, Shahrul Mt-Isa, Richard Nixon, and Deborah Ashby. A bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment. *Biometrical Journal*, 58(1):28–42, 2016.

Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

Yaping Wang, Yabing Mai, and Weili He. A quantitative approach for benefit-risk assessment using stochastic multi-criteria discriminatory method. *Statistics in Biopharmaceutical Research*, 8(4):373–378, 2016.

Grace Wangge, Olaf H Klungel, Kit CB Roes, Anthonius De Boer, Arno W Hoes, and Mirjam J Knol. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. *PLoS One*, 5(10):e13550, 2010.

Grace Wangge, Olaf H Klungel, Kit CB Roes, Anthonius de Boer, Arno W Hoes, and Mirjam J Knol. Should non-inferiority drug trials be banned altogether? *Drug discovery today*, 18(11-12):601–604, 2013.

Shihua Wen, Lanju Zhang, and Bo Yang. Two approaches to incorporate clinical data uncertainty into multiple criteria decision analysis for benefit-risk assessment of medicinal products. *Value in Health*, 17(5):619–628, 2014.

Brian L Wiens and Gerd K Rosenkranz. Missing data in noninferiority trials. *Statistics in Biopharmaceutical Research*, 5(4):383–393, 2013.

Brian L Wiens and William Zhao. The role of intention to treat in analysis of noninferiority studies. *Clinical Trials*, 4(3):286–291, 2007.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Bongin Yoo. Impact of missing data on type 1 error rates in non-inferiority trials. *Pharmaceutical statistics*, 9(2):87–99, 2010.